UNIVERSIDADE LUSÓFONA

UNIVERSIDADE LUSÓFONA

Proceedings of the

5th International Conference on AI Research

Hosted by Universidade Lusófona, Lisbon, Portugal

5-6 December 2024



Professor Carlos Gonçalves
Professor José Carlos Dias Rouco
Dr Carlos Rouco

A conference managed by ACI, UK

aci

Proceedings of the

5th International Conference on AI Research

ICAIR 2024

A Conference

Hosted By

Universidade Lusófona, Lisbon, Portugal

5-6 December 2024

Edited by

Professor Carlos Gonçalves Professor

José Carlos Dias Rouco

Copyright the authors, 2024. All Rights Reserved.

No reproduction, copy or transmission may be made without written permission from the individual authors.

Review Process
Papers submitted to this conference have been double-blind peer reviewed before final acceptance to the conference. Initially, abstracts were reviewed for relevance and accessibility and successful authors were invited to submit full papers. Many thanks to the reviewers who helped ensure the quality of all the submissions.

Ethics and Publication Malpractice Policy
ACIL adheres to a strict ethics and publication malpractice policy for all publications – details of which can be found here:
http://www.academic-conferences.org/policies/ethics-policy-for-publishing-in-the-conference-proceedings-of-academic-conferences-and-publishing-international-limited/

Self-Archiving and Paper Repositories
We actively encourage authors of papers in ACIL conference proceedings and journals to upload their published papers to university repositories and research bodies such as ResearchGate and Academic.edu. Full reference to the original publication should be provided.

Conference Proceedings
The Conference Proceedings is a book published with an ISBN and ISSN. The proceedings have been submitted to a number of accreditation, citation and indexing bodies including Clarivate Web of Science and Elsevier Scopus. Author affiliation details in these proceedings have bee reproduced as supplied by the authors themselves.

Author affiliation details in these proceedings have been reproduced as supplied by the authors themselves.

These proceedings are open access and freely available for all to read from https://papers.academic-conferences.org/index.php/icair. The Conference Proceedings for previous years can be purchased from http://academic-bookshop.com

E-Book ISBN: 978-1-917204-28-6
Print version ISBN: 978-1-917204-27-9
Print Version ISSN: 3049-5628

Published by Academic Conferences International Limited
Reading, UK
+44 (0) 118 324 6938
www.academic-conferences.org
info@academic-conferences.org

Contents

The papers in these proceedings were presented at the conference in sessions organised by topic, but are presented here by type of paper and then alphabetically by lead author surname.

About the Proceedings of the 2024 International Conference on AI Research

Preface **v**

Biographies **vii**

Review Committee **vi**

Academic Papers

The 4I Model of Benefits and Its Integration with the Technology Acceptance Model (TAM): Teenagers' Perspectives on Using ChatGPT for Homework
Zinaida Adelhardt, Thomas Eberle 1-9

Adopting Artificial Intelligence in Organisations: A Closer Look
Massimo Albanese 10-19

Generative AI Learning Environment for Non-Computer Science Engineering Students: Coding Versus Generative AI Prompting
Robert Alphinas, Torben Tambo 20-29

An Exploratory Study on the Role of Generative AI in Designing Educational Board Games
Fabrizio Amarilli, Oxana Timakova 30-39

Artificial Intelligence Life Cycle: The Detection and Mitigation of Bias
Ashionye Aninze 40-49

Designing an Artificial Intelligence Maturity Model for Human Resources (HR-AIMM)
Sascha Armutat, Malte Wattenberg, Nina Mauritz 50-58

Balancing AI in SMEs: Overcoming Psychological Barriers and Preserving Critical Thinking
Madeleine Block 59-66

The new era of Technology Mysticism: Generative Artificial Intelligence and its effects
Karsten Böhm, Jürgen Sammet 67-74

Artificial Intelligence and Trend Forecasting in the Fashion Industry:
Ethical and Anticipated Ethical Issues
Richard Wilson, Olivia Bowles 75-81

Advancing Corporate Social Responsibility in AI-driven Human Resources
Management: A Maturity Model Approach
Marta Bubicz, Marcos Ferasso 82-90

Bridging the Gap: Practical Challenges and Strategic
Imperatives in Adopting Gen AI
Andrea Di Vetta 91-98

The Academic Anti-Procrastination Approach: Combining Peer
Motivation and Personalized Artificial Intelligence Reminders
Xiaojiao Duan, Zhaoxia Yi,, Yongjia Sun, Itamar Shabtai 99-107

How Good Is GPT's "Emojinal Intelligence"? Investigating Emoji
Patterns in LLM-Generated Social Media Text
Michael Dunn, Kenneth Hopkinson 118-114

Online Managerial Tools as Research Tools to Apply Artificial
Management: Results of Research
Olaf Flak 115-125

Generative AI in Writing Workshops: A Path to AI Literacy
Sonja Gabriel 126-132

Generative AI and Educational (In)Equity
Sonja Gabriel 133-142

Exploring the Transformative Intersection of Artificial Intelligence
and Educational Research: K-12 Principals Supporting English Learners
Belinda Gimbert, Dustin Miller, Raeal Moore, Dean Cristol, Nick Giester 143-149

Citizens' Right to Information and the Principle of Good Administration:
What's new in the AI Act?
Anna Murphy Høgenhaug, Hanne Marie Motzfeldt 150-156

Artificial Intelligence and Neural Style Transfer in the Context of Art and Design:
Ethical and Anticipated Ethical Issues
Richard Wilson, Eunice Hong 157-164

Play my thesis: Opportunities and Limitations
of Building Serious Games with AI
Niklas Humble 165-174

Generative Artificial Intelligence and the Impact on Sustainability
Niklas Humble, Peter Mozelius 175-182

Unleashing Human Potential: A Framework for Augmenting
Co-Creation with Generative AI
Mitt Kabir 183-193

Artificial Intelligence in Banking : The Evidence from Poland
Monika Klimontowicz 194-202

Artificial Intelligence (AI) in the Dentist's Work: The Perspective
of Anthropomorphization and Cognitive Aspects of the Decision-Making Process
Guilherme Knupp Muniz, Luciana Paula Reis,
Sérgio Evangelista Silva, June Marques Fernandes 203-209

AI in the Learning Environment: Examination of Pedagogical,
Psychological, and Ethical Implications
Gala Krsmanovic, Fadi P. Deek 210-217

What The Phish! Effects of AI on Phishing Attacks and Defense
Shreyas Kumar, Anisha Menezes, Sarthak Giri, Srujan Kotikela 218-226

Governance Considerations of Adversarial Attacks on AI Systems
Nombulelo Lekota 227-233

Mapping Software-Engineering Industry AI Use to Software-
Engineering Curriculum: Developing the AI-USE Framework
Addison Lilholt, Thomas Heverin 234-242

Empowering Diversity by Building Inclusive Software Engineering
Projects with Large Language Models
Clara Maathuis, Greg Alpar, Stefano Bromuri 243-251

Navigating Online Narratives on Israel-Hamas War with LLMs
Clara Maathuis, Iddo Kerkhof 252-259

Unveiling the Emotional Landscape of Ukrainian War Narratives
using Large Language Models
Clara Maathuis, Iddo Kerkhof 260-270

AI, Personalized Education, and Challenges
Aziz Mimoudi 271-280

From Data to Decisions: Leveraging AI for Proactive Education Strategies
Willie Moore, Li-Shiang Tsay 281-288

Generative AI and its Impact on Activities and Assessment
in Higher Education: Some Recommendations from Master's Students
Peter Mozelius 289-295

Educating the Educators on Generative Artificial Intelligence
in Higher Education
Peter Mozelius, Marcia Håkansson Lindqvist,
Cleveland-Innes Martha, Jimmy Jaldemark, Marcus Sundgren 296-302

Transcending Fixed Meanings: Exploring the Impact of Linguistic
Relativism on Adaptive Language Models in Generative AI
Subhangi Namburi 303-312

GenAI Acceptance in Professional Services:
The Case of Management Consulting
Dennis Nesemeier 313-321

Ecosystem Theory and the Adoption of Artificial Intelligence in SMEs
Steve Nolan, Stelios Zyglidopoulos 322-328

Security and safety concerns in the age of AI
Victoria Yousra Ourzik 329-337

Ethical Generative AI: What Kind of AI Results are Desired by Society?
Marc Lehmann, René Peinl, Andreas Wagener 338-344

AI in Education: Balancing Innovation and Responsibility
Sruthi Pillai, R Ramakrishnan 345-354

Effective Continuous Quantitative Measures for End-to-End AI Guardrails
Jason Pittman 355-363

The Comparative Analysis of YOLOv5/v8/v9 for Object Detection,
Tracking, and Human Action Recognition in Combat Sports
Evan Quinn, Dr Niall Corcoran 364-373

AI Generative in Brazil’s Public Schools: The Teachers’ Perspective
Jussara Reis-Andersson 374-380

LLM Supply Chain Provenance: A Blockchain-based Approach
Shridhar Singh, Luke Vorster 390-397

Generative AI and the End of Education
Stockman 390-397

Can Robots Laugh? An Inquiry in the Nature of AI
Eduardo Tomé, Elizaveta Gromova 398-404

An Analysis of Knowledge Management Changes
Through Artificial Intelligence with Probst's Model
Burak Toptas 405-414

Artificial Intelligence and its Role in Legislative Practices
Marius Vacarelu 415-422

Artificial intelligence and the ethics of tomorrow
Brenda VAN WYK, Marlene Holmner 423-431

Comparing human-labeled and AI-labeled speech datasets for TTS
Johannes Wirth, René Peinl 432-438

PhD Papers

The next generation of AI tools are coming, do we need to better regulate them?
Michael Aubrey 440-447

“Should everyone have access to AI? "
Perspectives on Ownership of AI tools for Security
Yasmine Ezzeddine, Petra Saskia Bayerl 488-455

How Is Fake News Spread? An Analysis of the Dissemination Process:
Actors, Channels, and Motives
Anastasiia Iufereva 456-462

Towards the Virtual - AI Beauty Pageants and Their Implications for Society
Ida Knapik 463-471

Masters Papers

Exploring Student Perspectives on Generative AI in Requirements
Engineering Education
Nicklas Mellqvist, Peter Mozelius 473-481

Late Submissions

Challenges in AI Implementation: Perspectives from Practice and Research
Clemens Kerschbaum, Raphael Dachs 483-490

Driving the Future: AI in Transportation
Isabella Xu, Murtaza Haque, Song Fu 491-499

Students' AI-generated Images: Impact on Motivation, Learning and, Satisfaction
Cornelia Berg, Liv Omsén, Henrik Hansson, Peter Mozelius 500-506

About the 5th International Conference on AI Research, 2024

The International Conference on AI Research (ICAIR) is becoming a key forum for academics and students, whether social scientists, natural scientists or philosophers come together to explore the future of society in the context of AI. Many aspects of AI are now the subject of peer reviewed academic research, including the increasing use of Chatbots and GenAI in academe and in society at large.

The conference was initially launched in 2019 as the European Conference on Artificial Intelligence and Robotics. After a success inaugural event in Oxford the conference fell victim to Covid and the following three years where held online. Time was taken in 2023 to refocus this fast developing area of research and the name of the conference was changed this year to the International Conference on Artificial Intelligence Research.

The Proceedings of the 2024 edition of the conference include academic research papers, PhD research papers, Masters Research papers and work-in-progress papers, all of which have been presented and discussed at the ICAIR 2024 conference. The proceedings are of an academic level appropriate to a professional research audience, including graduates, post-graduations, doctoral and post-doctoral researchers. All papers have been double-blind peer reviewed by members of the Review Committee.

Aims and Scope

The International Conference on AI Research (ICAIR) is an opportunity for academics, practitioners, and consultants from around the world who are involved in the study, management and development of artificial intelligence related issues in education, in business or in the public sector to publish and present their research. It seeks to bridge the gap between academic research, practical applications and policy-making. With a multi and trans-disciplinary scope the conference attracts research from diverse settings.

Topics Covered

The call for papers for ICAIR 2024 asked for contributions that considered the following topics. In addition the committee welcomed papers on a number of specialist mini-tracks which can be seen at the end of this list.

- The consequences of AI on work and society
- The governance of AI and AI in governance
- Regulations for AI safety
- Enabling knowledge networks with cognitive computing and other technologies
- Ethical and acceptable social behaviours
- Organisational and social impact of AI

- How can AI explain its thinking?

Further applications of AI including:

- GenAI and AI Chatbots
- Fintech and the financial services industry
- Life sciences and healthcare
- Utilities and transport management services
- In the development of Smart-cities
- In education and training
- In scientific research
- In healthcare
- Collaborative robots and human-robot interaction
- Society 5.0

Experts in the field proposed mini tracks on the following topics. Papers accepted on these topics after the double-blind peer review process were presented as mini tracks at the conference and are also published in these proceedings.

Mini Tracks

- Transforming the Future: How Generative AI and LLMs are Revolutionizing Digital Transformation
- Ethical and Societal Implications of AI
- AI and Educational Research and Researchers
- Examining the Future of AI and Human Systems Integration

The programme for the 2024 conference can be downloaded from here and full details about the conference can be found on the conference website:

https://www.academic-conferences.org/conferences/icair/

ICAIR Preface

These proceedings represent the work of contributors to the 2024 International Conference on AI Research (ICAIR) hosted by ACI and the Universidade Lusófona, Lisbon, Portugal on 5-6 December 2024. The Conference Chair was Professor Carlos Gonçalves and the Programme Chair was Professor José Carlos Dias Rouco both from Universidade Lusófona.

ICAIR is fast being recognised as well-established event on the academic research calendar and now in its key aim remains to provide the opportunity for participants to share ideas and meet the people who hold them. The scope of papers accepted for the conference this year provided participants with a most interesting two days. The subjects covered illustrated the wide range of topics that fall into this important and fast-growing area of research.

The opening keynote presentation was given by Professor José Borges from the Portuguese Military Academy, who will talk about The Impact of AI Limitations in Relation to Defence Applications. The second day of the conference was opened with an address from Carlos Maio, Executive Manager and Founder QSR who discussed AI: The Demographic Inverted Pyramid.

With an initial submission of 127 abstracts, after the double blind, peer review process there are 54 Academic research papers, 4 PhD research papers, and 1 Masters Research papers published in these Conference Proceedings. These papers represent research from Brazil, Canada, Denmark, France, Germany, Italy, PERÚ, Poland, Sweden, UK, and USA.

We would like to thank everyone who participated in this conference.

Carlos Gonçalves and José Carlos Dias Rouco
Universidade Lusófona, Lisbon, Portugal
December 2024

Biographies of Chairs and Keynote Speakers

Conference and Programme Chairs



Carlos Gonçalves is an Associate Professor at Lusophone University - School of Economic Sciences and Organizations - Department of Management on Civil Aviation and Airports. Researcher on Complexity Sciences, Chaos Theory, Risk Science, Quantum Technologies, Artificial Intelligence, Strategic Studies, Studies in Intelligence and Security, FinTech and Financial Risk. Software developer in the fields of Complexity Sciences, Chaos Theory, Artificial Intelligence, Data Science and Quantum Technologies. Digital artist working on fractal art, music and applications of generative AI to both visual art and music



José Carlos Dias Rouco is an Associate Professor of Human Resource Management in the School of Economic Sciences and Organizations at Lusofóna University. He obtained her PhD from Lusíada University in 2012. He holds a Master's degree in Sports Training Management from Faculty of Human Motricity at Technical University of Lisbon. He has a graduation (5 years) in Military Sciences from the Portuguese Military Academy. Currently, he is the director of the Department of Air Transport and Airports Management, Director of the Bachelor's degree in Aeronautical Management. He teaches courses on crisis management, strategic and leadership, leadership and ethics, strategic marketing, international negotiations, human resource management, development and vocational training and organizational behavior. He has taught and coordinated several intensive leadership courses (internal regime for 5 days) for young university students, entrepreneurs, bankers, and firefighters. Currently, he is a researcher of a research project on Stressense funded by FCT, which was awarded the scientific prize. He is a regular reviewer for many International journals.

Keynote Speakers



José Borges is a mechanical engineer specialising in systems, automation, control and robotics. He has a PhD from Instituto Superior Técnico in nonlinear system modelling. José worked as a professor and researcher at Instituto Superior Técnico and at the Delft Centre for Systems and Control in the Netherlands. He has specialised education in Cyber Security from the Massachusetts Institute of Technology, Management from the Lisbon School of Economics – ISEG, and Management and Innovation, Entrepreneurship and Technologies by COTEC Portugal, FLAD and the University Institute of Lisbon – ISCTE. José holds a teaching and research position with the Portuguese Military Academy, where he has been the Executive Director of the Military Academy Research Centre – CINAMIL, the Scientific Coordinator of the Military Mechanical Engineering courses and Director of the Mechanical Engineering Laboratory. He is also Deputy President of the AFCEA Portuguese chapter. His current R&D projects, funded by Horizon Europe and European Defence Fund programmes, include cyber defence, cybersecurity, artificial intelligence, and command and control systems applications.



Carlos Maio has more than 17 years of experience in Human Resources Consulting and Business Management, during this time he has focused his skills on the Aerospace and Defence sectors and more recently, on Maritime and Energy industries. Since 2014, Carlos has been the Executive Manager and founder of QSR, a strategy consulting company developing high-end solutions within human capital, specialising in industries with high technological intensity. QSR was granted the leading role of the EACP (European Aerospace Cluster Partnership) Skills Working Group whose main objective is to integrate the best practices of Human Capital sciences in the aerospace SME's, enhancing the capacity of the sector to integrate disruptive and high productivity technologies, products and processes. Carlos has also lead the Skills WG of the Portuguese Aerospace Cluster and Portuguese Renewable Energy Cluster. He has been a

member of Pact for Skills for Aerospace and Energy involved in various European R&D projects for Skills in Aerospace and Maritime. Recently, Carlos has launched a QSR Spin Off, which will aim to use ML and AI to boost effectiveness and efficiency in Skills Management. This platform will be useful for Public Entities, clusters, training entities, companies and individuals, giving them information and knowledge about the tendencies of demand and offer of competencies in the market.

Mini Track Chairs



Lori D. Coombs, MBA, MSE is an Assoc Prof at the Univ of Arizona Global Campus. Her career began with former US Vice Pres Candidate & Senator Joe Lieberman. She was appointed Delegate by Joe & former US Vice Pres Al Gore in 2000. She is CEO of WWCM, & Director of WWCM Academy, a nonprofit. She supports teams at NASA & C6ISR DoD programs. She is pursuing a D.Sc. in Cybersecurity at Marymount Univ. Her work on, "AI Risks in Higher Ed," is published in the EJEL. Research: Cyberpsychology, Geology, Astronomy, Heliophysics, Space Policy, Intl Space Law, Title 51USC.



Dean Cristol, Ph.D., associate professor, Department of Teaching & Learning, The Ohio State University, Columbus, OH focuses on educational technology professional development within marginalized educational settings, aiming to equip both students and teachers with culturally responsive skills for technology-driven learning environments. Engages in research initiatives funded by international, national, state, and private foundation grants. Contributes extensively to academic discourse through publications in numerous journals, authoring books and chapters, and presenting findings at international, national, and state conferences. He serves as an Associate Editor for the Theory Into Practice journal and is Vice President for the International Association for Mobile Learning.



Nick Giester, MBA, more than 20 years of experience in public and private sector industries of IT, specializing in delivering technological solutions to business partners focused on the automation and improvement of vital business operations, as well as adeptly navigating organizational culture. He is highly skilled in evaluating the compatibility of existing infrastructure and software with fundamental business services, offering strategic guidance on investing in technologies like artificial intelligence or deciding on divestitures to meet business objectives. He excels at building and sustaining critical relationships, enabling him to manage crises efficiently, troubleshoot issues, and broker solutions effectively.



Belinda G. Gimbert, Ph.D., associate professor, Educational Administration, Department of Educational Studies, The Ohio State University, Columbus, OH. Her research addresses a broad array of human capital policies that impact talent management in low performing K-12 school systems and influence the composition, distribution, and quality of educators in the workforce. Gimbert teaches courses related to human resource administration, introduction to educational administration, and K-12 instructional supervision focusing on using AI and mobile learning technologies. She also directs two national projects sponsored by the U.S. Department of Education's Office of English Language Acquisition.



Dr. Mitt Nowshade Kabir is an adjunct professor of Digital Transformation at Ecole de Management Léonard De Vinci. He is a serial entrepreneur, knowledge evangelist, business accelerator, and an investor. Dr. Kabir has a master's degree in computer science, an MBA, a Ph.D. in Information Technology, and a DBA in Management. His current interests include AI, innovation, entrepreneurship, and strategic management.



Dustin Miller, Ph.D., clinical assistant professor, Educational Administration, Department of Educational Studies, The Ohio State University, Columbus, OH. His research interests focus on principal professional learning, leading in times of crises, and creating supportive LGBTQ+ environments for school leaders and teachers. He teaches courses on leadership, professional learning, and human resource management. Miller spent 20 years in k-12 as a high school principal, middle school principal, and director of secondary education. He currently serves as the plenum representative for the University Council for Educational Administration. He also served as president of the Ohio Association of Secondary School Administrators.



Raeal Moore, Ph.D., Executive Director of Evaluation Data Solutions, LLC. She specializes in providing comprehensive, high-quality research, evaluation, and assessment services to deliver reliable and accurate information about the implementation and outcomes of programs and projects. She uses mixed methods approaches to answer programmatic questions. She understands that a credible and comprehensive evaluation is critically important to provide program decision-makers with timely, evaluation results. With 20 years' experience, Raeal spends time strengthening the evaluation capacity and practice of organizations. Moore earned an M.A. and Ph.D. in Quantitative Research, Evaluation, and Measurement at Ohio State University.

ICAIR Conference Committee

ICAIR is fortunate to have a significant number of international experts in the field willing to review papers. Care is taken to ensure that a fully double-blind peer review process is followed.

Dr Kareem Kamal A.Ghany, Beni-Suef University, Egypt; Prof Azween Abdullah, Taylors University, Malaysia; Dr Kinaz Al Aytouni, Arab International University, Syria; Dr Hanadi AL-Mubaraki, Kuwait University, Kuwait; Prof Hamid Alasadi, Iraq University College, Iraq; Dr Saad Aldoihi, KACST, Saudi Arabia; Prof Laurice Alexandre, Sorbonne Paris Cité , France; Dr José Álvarez-García, University of Extremadura, Spain; Dr Xiomi An, Renmin University of China, China; Prof Antonios Andreatos, Hellenic Air Force Academy, Greece; Prof Oscar Arias Londono, Institucion Universitaria de Envigado, Colombia; Dr Gil Ariely, Interdisciplinary Center, Herzliya, Israel; Prof Atef Ata, Faculty of Engineering, Alexandria University, Egypt; Dr Sotiris Avgousti, Cyprus University Of Technology, Cyprus; Prof Rosalina Babo, Polytechnic of Porto, Porto Accounting and Business School, Portugal; Prof Liz Bacon, Abertay University, UK; Prof Jorge Barbosa, ISEC/IPC, Coimbra Institute of Engineering/Polytechnic Institute of Coimbra, Portugal; Prof Jebin Bose S, Noorul Islam Centre for Higher Education, Kumaracoil, India; Prof Larbi Boubchir, University of Paris 8, France; Assc Adrian Nicolae Branga, Lucian Blaga University of Sibiu, Romania; Dr Ersin Çağlar, European University of Lefke, Lefke / TRNC / Turkey; Prof Karen Cham, University of Brighton, UK; Dr Divya Chaudhary, Northeastern University, USA; Prof Jim Chen, U.S. National Defense University, USA; Dr Pericles Cheng, European University Cyprus, Cyprus; Prof Koteshwar Chirumalla, Malardalen University, Sweden; Mr David Comiskey, Ulster University, UK; Assoc.Prof.Lori D. Coombs, Marymount University, USA; Dr Leonardo Costa, Universidade Católica Portuguesa - Católica Porto Business School, Portugal; Dr Dean Cristol, The Ohio State University, USA; Prof Larry Crockett, Augsburg University, USA; Dr Marija Cubric, University of Hertfordshire, UK; Francesca Dal Mas, Università di Udine/Università la Sapienza, Italy; Assc Ben Daniel, University of Otago, New Zealand; Prof Justine Daramola, Cape Peninsula University of Technology, South Africa; Geoffrey Darnton, WMG, University of Warwick, UK; Mr Martin De Bonis, Alma Mater Studiorum,

Italy; Prof Armando Carlos de Pina Filho, Federal University of Rio de Janeiro (UFRJ), Brazil; Dr Martin De Saulles, University of Brighton, UK; Dr María de la Cruz Del Río-Rama, University of Vigo, Spain; Dr Souâd Demigha, CRI Univ of Paris 1 La Sorbonne, France; Paolo Di Muro, Politecnico di Milano School of Management, Italy; Dr Mihaela Diaconu, "Gheorghe Asachi" University of Iasi, Romania; Inês Domingues, DEIS-ISEC, Portugal; Dr Patricio Domingues, Polytechnic Institute of Leiria, Portugal; Prof Yanqing Duan, University of Bedfordshire, UK; Prof. John Edwards, Aston Business School, UK; Dr Kelechi Ekuma, The Global Development Institute, University of Manchester, UK; Dr Scott Erickson, Ithaca College- School of Business, USA; Dr José Esteves, IE Business School, Spain; Prof Rui da Assunção Esteves Pimenta, Politécnico do Porto, Portugal; PhD Fernanda Faini, CIRSFID - University of Bologna, Italy; Dr Georgios Fessakis, University of the Aegean, Greece; Prof Eric Filiol, ENSIBS, Vannes, France & CNAM, Paris, France; Dr Panagiotis Fotaris, University of Brighton, UK; Prof Andreas Giannakoulopoulos, Ionian University, Greece; Dr Belinda Gimbert, The Ohio State University, USA; Dr Valerie Priscilla Goby, Zayed University, UAE; Dr Amol Gore, UAE Government HCT, UAE; Dr Paul Griffiths, Ecole de Management de Normandie, Oxford, UK; Dr Hossein hakimpour, IAU , Iran; Prof Ali Hessami, City University London, UK; Dr Grant Howard, University of South Africa (Unisa), South Africa; Prof Hamid Jahankhani, Northumbria University London, UK; Dr Surbhi Jain , Baba Mastnath University Rohtak , India; Assc Jimmy Jaldemark, Mid Sweden University, Sweden; Dr Aman Jatain, Amity University, India; Dr Jari Jussila, Häme University of Applied Sciences, Finland; Dr Selvi Kannan, Victoria University, Australia; Prof Ergina Kavallieratou, University of the Aegean, Greece; Prof Thomas Keller, Institute of Business Information Technology, Switzerland; Dr Harri Ketamo, Headai ltd, Finland; Dr Nasrullah Khilji, University of Bedfordshire, UK; Prof Tatiana Khvatova, Emlyon Business School, Ecully, France; Prof Jesuk Ko, Universidad Mayor de San Andres, Bolivia; Prof Michael Kohlegger, Institute for Web Technologies & Applications, Austria; Prof Renata Korsakiene, Vilnius Gediminas Technical University, Lithuania; Prof Ibrahim Krasniqi, University of Peja, Kosovo; Prof Konstadinos Kutsikos, Business School, University of the Aegean, Greece; Dr Jean Lai, Hong Kong Baptist University, Hong Kong; Dr Isah Abdullahi Lawal, Noroff University College, Norway; Prof Fuhua (Oscar) Lin, Athabasca University, Canada; Dr Efstratios Livanis, University of Macedonia, Greece; Prof Joao Alexandre Lobo Marques, University of Saint Joseph, Macau; Dr Kartheesan Logu, Annamacharya Institute of Technology and Science, India; Prof Eurico Lopes, Instituto Politécnico de Castelo Branco, Portugal; Dr Martin Magdin, Constantine the Philosopher University in Nitra, Faculty of Natural Sciences, Slovakia; Dr Paolo Magrassi, Alephuture, Switzerland; Dr Juan Jose Martinez Castillo, Acantelys Research Group, China; Prof António Martins, Universidade Aberta, Portugal; Dr Panicos Masouras, Cyprus University of Technology, Cyprus; Prof Maurizio Massaro, Ca' Foscari University of Venice, Italy; Dr Nuno Melão, Polytechnic Institute of Viseu, Portugal; Dr Hacene Mellah, Bouira University, Algeria; Prof Anabela Mesquita, Polytechnic of Porto, Portugal; Prof David Methé, Institute of International Strategy at Tokyo International University in Tokyo, Japan; Prof Awais Mian, LUMS, Pakistan; Dr Larisa Mihoreanu, ANMDM Bucharest, Romania; Dr Clemente Minonne, Lucerne University of Applied Sciences, Institute for Innovation and Technology Management, Switzerland; Dr Nikolaos Misirlis , HAN University of Applied Sciences , The Netherlands; Prof Harekrishna Misra, Institute of Rural Management Anand, India; Assc Ludmila Mladkova, University of Economics Prague, Faculty of Business Administration, Czech Republic; Dr Artur Modliński, University of Łódź - Faculty of Management, Poland; Prof Fernando Moreira, Universidade Portucalense, Portugal; Prof John Morison, Queen's University Belfast , UK; Dr Rabeh Morrar, Northumbria University, UK;

Dr Hafizi Muhamad Ali, Yanbu University College, Saudi Arabia; Dr Antonio Muñoz, Universidad de Málaga, Spain; Prof Mihaela Muresan, Dimitrie Cantemir Christian University, Romania; Dr Minoru Nakayama, Tokyo Institute of Technology, Japan; Dr MItt Nowshade Kabir, Trouvus, USA; Assoc. Prof. Dr. Nuran Öze, Arkın University of Creative Arts and Design (ARUCAD), Northern Cyprus; Dr Pankaj Pandey, Norwegian University of Science and Technology (NTNU), Norway; Dr Jaydeep Patil, AISSMS IOIT Pune, India; Assc Corina Pelau, Bucharest University of Economic Studies,, Romania; Dr Parag Pendharkar, Penn State Harrisburg , USA; Dr Alexander Pfeiffer, Applied Game Studies, Austria; Dr Cosmin Popa, Innovative Agricultural Services, UK; Prof Ricardo Queirós, ESMAD/P.Porto, Portugal; Prof Carlos Rabadão, Politechnic of Leiria, Portugal; Dr Jaichandran Ravichandran, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation, India; Assc Liana Razmerita, Copenhagen Business School, Denmark; Dr Jussara Reis-Andersson, Department of Education, Sweden; Dr Marcin Relich, University of Zielona Gora, Poland; Prof José Carlos Ribeiro, Polytechnic Institute of Leiria, Portugal; Prof Sandra Ribeiro, ISCAP, IPP-Porto, Portugal; Dr Martin Rich, Cass Business School, UK; Dr Kenneth Rogerson, Sanford School of Public Policy, USA; Prof Göran Roos, University of South Australia, Australia; Dr Eleni Rossiou, Experimental School of Aristotle University , Greece; Prof Neil Rowe, U.S. Naval Postgraduate School, USA; Dr Melissa Saadoun, EIVP, France; Prof Lili Saghafi, MSMU, USA; Prof Mustafa Sagsan, Cyprus Inter Univ, Turkish Rebuplic of Northern Cyprus; Prof Abdel-Badeeh Salem, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt; Dr Char Sample, US Army Research Laboratory, USA; Dr Navjot Sandhu, Birmingham City University, UK; Prof Vitor Santos, NOVA IMS - New University of Lisbon, Portugla; Prof Ramanamurthy Saripalli, Pragati Engineering College, India; Prof Markus Schatten, Faculty of Organization and Informatics, University of Zagreb, Croatia; Assistant Professor Sandro Serpa, Universidade dos Açores, Portugal; Dr Armin Shams, Sharif University of Technology, Iran; Dr Yilun Shang, Northumbria University, UK; Dr Eric Shiu, University of Birmingham, UK; Prof Fernando Silva, Polytechnic Institute of Leiria, Portugal; Prof Andrzej Sobczak, Warsaw School of Economics, Poland; Dr Caroline Stockman, University of Winchester, UK; Dr Darijus Strasunskas, Hemit, Norway; Dr Olga Strikulienė, Kaunas University of Technology, Lithuania; Dr Marta-Christina Suciu, Bucharest University of Economic Studies, Romania; Prof Ramayah Thurasamy, Universiti Sains Malaysia, Malaysia; Assc Milan Todorovic, Royal Melbourne Institute of Technology, Australia; Prof Jim Torresen, University of Oslo, Norway; Dr khan ferdous wahid, Airbus, Germany; Prof Fang Wang, Nankai University, China; Prof Murdoch Watney, University of Johannesburg, South Africa; Prof Bruce Watson, Stellenbosch University, South Africa; Dr Santoso Wibowo, Central Queensland University, Australia; Prof Robert J. Wierzbicki, University of Applied Sciences Mittweida, Germany; Dr Marcus Winter, University of Brighton, UK; Mr Jason Wong, KDU University College, Malaysia; Mr Tuan Yu, Kent Business School, UK; Prof Daiva Zostautiene, Kaunas University of Technology, Lithuania; Dr Justyna Żywiołek, Czestochowa University of Technology, Poland; Prof João M. Lopes, Instituto Superior Miguel Torga, Coimbra, Portugal & NECE-UBI – Research Unit in Business Sciences, University of Beira Interior, Covilhã, Portugal, Portugal; Dr Nomalanga Mashinini, University of the Witwatersrand, South Africa; Dr William Joseph Spring, University of Hertfordshire, UK; Prof Nuno Domingues, ISEL, Portugal; Prof Sofia Gomes, Universidade Portucalense, Portugal; Prasad Kodaypak, TheXMG, USA; Dr Marta Bubicz, Lusófona University, Portugal; Dr Mourya Reddy Narasareddygari, Rider University, United States.

The 4I Model of Benefits and its Integration with the Technology Acceptance Model (TAM): Teenagers' Perspectives on Using ChatGPT for Homework

Zinaida Adelhardt and Thomas Eberle
Friedrich-Alexander University Erlangen-Nürnberg, Germany

zinaida.adelhardt@fau.de
office@thomaseberle.de

Abstract: The introduction of ChatGPT has provided students with a tool similar to a fairy-tale djinn, giving the ability to simplify homework and projects with just a few keystrokes. This study examines teenagers' perspectives on using ChatGPT for homework, exploring both the perceived benefits and concerns. Based on content analysis of responses from 141 teenagers aged 14 to 16 from grammar schools across various federal regions in Germany, we introduce the *4I Model of Benefits*, identifying four key advantages: Information, Inspiration, Improvement, and Immediacy. Integrated within the Technology Acceptance Model (TAM) framework, this model clarifies factors influencing teenagers' acceptance of AI text generators in education. Findings indicate that, while most teenagers recognize the distinct benefits of using ChatGPT, concerns about learning efficacy and over-reliance underscore the need for balanced educational integration. By enriching the TAM framework, the 4I Model identifies the specific benefits of ChatGPT that drive its acceptance and provides valuable insights for educators, policymakers, and developers.

Keywords: ChatGPT in education, 4I model of AI benefits, Educational benefits of AI text generators, Homework support through AI, Technology acceptance model (TAM) Extension, User studies

1. Introduction

The rapid growth of artificial intelligence (AI) has significantly impacted the educational landscape, offering students access to innovative easy-to-use tools having the potential to reshape and revolutionize learning experiences (Bem Machado et al. 2024; DuBoulay et al. 2023; Holmes et al. 2019). Among these tools are generative pre-trained transformers (GPTs) like OpenAI's ChatGPT, which stand out as particularly impactful in educational settings, surpassing other similar generative transformers such as Bing Chat, Bard, and Ernie (Rudolph et al. 2024). Like a fairy-tale djinn, ChatGPT grants students the ability to generate human-like responses, reasonable explanations, and even complete essays with just a few keystrokes – the powerful ability of new technology that raises both enthusiasm and concern among educators, students, and researchers.

In this article, we introduce the **4I Model of Benefits**—Information, Inspiration, Improvement, and Immediacy—developed from a qualitative analysis of 141 essays written by teenagers reflecting on their use of ChatGPT. The model identifies key advantages students perceive in using ChatGPT for their schoolwork. Furthermore, we situate these findings within the framework of the Technology Acceptance Model (TAM) (Davis 1989), a widely used framework that explain how users come to adopt and engage with new technologies. By combining the 4I Model with TAM, we aim to provide deeper insights into the factors that influence teenagers' acceptance of AI-based textual generators, highlighting the perceived benefits of ChatGPT among students.

1.1 ChatGPT in Education

The academic discourse surrounding ChatGPT's integration into education has emphasized its transformative potential in reshaping how students learn and how educators teach (Bem Machado et al. 2024; Farrokhnia et al. 2024; Grassini 2023; Tlili et al. 2023; Meyer et al. 2023). Most studies have focused on university settings, exploring its applications in specific subject areas, such as legal education (Choi et al. 2023), medical education (Kung et al. 2023; Wen and Wang 2023), media education (Pavlik 2023), and programming (Rahman and Watanobe 2023). ChatGPT's relatively high performance on the US Medical Licensing Exam (Kung et al. 2023) and its ability to handle law school exams at an average student level (Choi et al. 2023) highlight its potential to assist in higher education contexts.

1.2 ChatGPT and its Benefits for Students

OpenAI's ChatGPT has gained considerable attention for its potential to assist students with a wide array of academic tasks (Grassini 2023; Mahapatra 2024; Meyer et al. 2023; Sok and Heng 2023). Trust et al. (2023) outlined several key benefits of ChatGPT for students, starting with *personalized learning support*, which includes individualized tutoring, research assistance, activity directions, and explanations of complex topics in simpler language. It can generate summaries, outlines, writing tips, sentence starters, and make translations

(Trust et al. 2023). Moreover, ChatGPT *supports creative thinking* by suggesting ways to improve sentence structure and aiding brainstorming processes. It also provides *assessment support*, such as quizzes and practice tests, and assist *reading and writing comprehension*. Existing studies affirm these benefits, demonstrating ChatGPT's ability to improve academic writing skills (Mahapatra 2024), offer personalized tutoring, and assist with various research tasks (Grassini 2023).

1.3 Concerns about ChatGPT

Despite its many benefits, the use of ChatGPT in education has sparked significant concerns, particularly around *academic integrity* (Kasneci et al. 2023; Trust et al. 2023; Tlili et al. 2023; Sok and Heng 2023). The ease with which it generates human-like text has led to fears regarding the preservation of academic honesty and its potential misuse in educational settings (Cotton et al. 2023; Sullivan et al. 2023).

Additionally, ChatGPT has been found to provide *inaccurate information*, fabricate content, create false citations, and generate unreliable reference lists, all of which compromise its credibility (Azaria et al. 2024; Bhattacharyya et al. 2023). Research also highlights risks of reproducing specific negative perceptions (Abramski et al. 2023) and political biases (Rozado 2023).

Over-reliance on AI poses another critical challenge, with concerns that tools like ChatGPT may erode students' higher-order cognitive skills and reduce independent problem-solving abilities (Farrokhnia et al. 2024). Uncritical engagement with AI-generated content also raises issues for educators and policymakers, alongside concerns about copyright, data protection, legal responsibility, and cybersecurity in education (Sallam 2023).

These concerns are driving ongoing discussions about the responsible use of ChatGPT and debates on whether it should be allowed in educational contexts (Yu 2023; Ogugua et al. 2023). While several Australian states have restricted the use of ChatGPT in schools to maintain educational standards (Bernardone 2023), some US schools are experimenting with generative AI tools as a potential substitute for teachers (Young 2023). However, even if schools ban ChatGPT on school devices, controlling how students complete their homework remains challenging. Current detection tools for identifying AI-generated content have proven to be ineffective (Pegoraro et al. 2023). This underscores the need for more research on the ethical integration of ChatGPT in education (Cotton et al. 2023; Miao et al. 2021; Hagendorff 2020; Nguyen et al. 2023). However, in order to develop reliable ethical policies and guidelines, it is essential to further investigate how students use these tools and what drives their appeal.

1.4 Theoretical Framework and Research Objectives

The use of ChatGPT in education has predominantly been studied within the context of higher education, with a significant focus on university students from various disciplines (Bašić et al. 2023; Chan and Tsi 2023; Al-Abdullatif and Alsubaie 2024; Ravšelj et al. 2024), leaving a critical gap in understanding how younger, more vulnerable students in secondary schools engage with these tools.

Research on secondary school students' use of ChatGPT is limited, despite teenagers being key users of the technology. OpenAI's Terms of Use require users to be at least 13 years old (OpenAI.com 2024), and with growing access among teens, it is crucial to understand their views on its benefits and risks. Unlike university students, secondary school students face different cognitive and developmental challenges, highlighting the need for tailored policies and guidelines for them. More research is needed on how younger students use ChatGPT, its appeal, and the ethical implications in secondary education.

To address this gap, our study investigates how secondary school students use ChatGPT, with a focus on the benefits they perceive in relation to their schoolwork. We aim to explore the key factors that make ChatGPT appealing to teenagers, particularly in how they believe it enhances their learning experience. The central research question guiding this study is: "What benefits do teenagers perceive in using AI text generators like ChatGPT for their schoolwork?"

To analyze the factors influencing teenagers' acceptance of ChatGPT, we apply the **Technology Acceptance Model (TAM)**, originally developed by Davis (1989). TAM posits that two main factors influence users' acceptance of new technologies:

- **Perceived Usefulness (PU)**: The belief that using the technology will improve one's performance.
- **Perceived Ease of Use (PEOU)**: The belief that using the technology will be effort-free.

These factors are influenced by external variables and shape users' attitudes toward the technology, which in turn influence their behavioural intentions and ultimately determine actual usage. While TAM has been

extended in various ways, including TAM 2 (Venkatesh and Davis 2000) and the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al. 2003), we choose to use the original TAM framework for its simplicity. However, we extend it by introducing an additional layer to capture the specific appeal of generative AI tools like ChatGPT. Bagozzi (2007) criticized TAM for potentially oversimplifying the complexity of technology adoption, particularly across different contexts and user groups, stating that "almost no research has deepened TAM in the sense of explaining PU and PEU, reconceptualizing existing variables, or introducing new ones to explain how these variables produce their effects" (p. 244). Our approach aims to address this gap by adapting TAM to the context of generative AI, specifically focusing on ChatGPT adoption.

This study introduces an extension of TAM through the development of the **4I Model of Benefits**—Information, Inspiration, Improvement, and Immediacy. By integrating this model within the TAM framework, we aim to offer a more nuanced understanding of the factors influencing ChatGPT's acceptance among secondary school students. While prior research has explored the use of generative AI tools like ChatGPT in education, few studies have focused specifically on teenagers' perceptions. This study presents the 4I Model, derived inductively from qualitative data, as a novel approach to understanding the perceived benefits of text generators like ChatGPT among school students.

In doing so, we seek to contribute to ongoing discussions about AI generators in education by providing evidence-based recommendations for ethical integration. By understanding what drives ChatGPT's appeal for younger students, we aim to inform the development of policies and guidelines that promote responsible, effective use of AI generative tools in schools, ensuring it supports rather than hinders educational outcomes.

2. Methodology

This study employs a qualitative approach to examine teenagers' perceptions of the benefits associated with AI text generators. A total of 141 teenagers, aged 14 to 16 (mean age 14.97, standard deviation 0.77, 58.7% female), participated in the research. The participants, students from German grammar schools (Gymnasien) and applicants to an adventure education program, represented a geographically diverse sample, coming from fourteen of Germany's sixteen federal regions.

Participants were asked to write short essays in response to the question: "What advantages do you see in using ChatGPT as a resource for homework or projects?" Given the involvement of participants under the age of eighteen, we implemented several safeguarding measures. Written informed consent was obtained from both participants and their parents/guardians. To ensure clarity, we conducted two pre-data collection Zoom meetings to explain the research aims and methods. All responses were anonymized to protect participants' identities, stored securely, and accessible only to the research team. Participation was voluntary, with the option to withdraw at any time. The survey question was carefully designed to be age-appropriate and understandable for teenagers.

Data collection occurred via the online survey platform SoSci Survey at the end of February 2024. Participants were given ten days to complete the survey at their convenience, allowing them reasonable time to provide thoughtful and well-considered responses. The length of responses to the open-ended question varied from one to fifty-six words.

A content analysis was performed on the collected data, categorizing the responses into themes based on perceived benefits. Preliminary thematic analysis revealed four key categories of benefits - Information, Inspiration, Improvement, and Immediacy - which led to the development of the 4I Model. Unlike other existing mnemonic-based frameworks, the 4I model was derived directly from participants' feedback, specifically reflecting their views on the benefits of ChatGPT. Additionally, responses highlighting a lack of perceived benefits or focusing on drawbacks were also identified and quantified as part of the analysis.

3. Results

The analysis of the teenagers' essays revealed that all participants were already familiar with ChatGPT and held well-defined opinions about its usefulness for academic work. Notably, 83% of teenagers identified specific benefits that they perceived in using ChatGPT, while 17% reported that they saw no advantages in this tool.

3.1 Perceived Benefits of ChatGPT

The analysis of teenagers' essays highlighting benefits revealed a consistent set of recurring themes - **Information, Inspiration, Improvement, and Immediacy** - that capture the perceived advantages of using ChatGPT for academic tasks. All benefits mentioned by students can be aligned with one of these categories.

The distribution of responses across these themes suggests a varied range of uses for AI text generators, emphasizing both practical and creative aspects. Below are the benefits that teenagers associate with using ChatGPT (the total percentage exceeds 83% because many students mentioned multiple benefits in their responses):

3.1.1 Information

A significant portion of respondents (27%) highlighted the value of ChatGPT in providing quick access to vast amounts of information. Teenagers appreciated how ChatGPT could condense research time by offering readily available summaries and explanations from diverse sources. For many students, this was particularly useful when they needed to grasp complex topics or needed a starting point for further investigation. The ability to retrieve relevant content instantly is described as helpful in understanding subjects and in creating a more holistic view of the subject matter. One student noted an additional advantage in using ChatGPT over general internet search engines: it provides information without the distraction of unnecessary advertisements.

> *Examples: "You can gather a lot of different information from different sources quickly and get an overall picture" (P.22); "Questions can be answered appropriately, by an AI pulling information from the internet and adapting it to the question" (P.57); "You can search more specifically and find content faster. Plus, you don't get unnecessary info/ads" (P.116).*

3.1.2 Inspiration

For many teenagers, ChatGPT functions as a creative spark (16%). The tool's ability to present ideas in novel ways or suggest alternative approaches made it valuable for students seeking inspiration. It was often described as a way to break through creative blocks, particularly when beginning an assignment or project. By offering fresh perspectives, ChatGPT served as a catalyst for generating new ideas, enhancing students' ability to explore different solutions or angles on a task.

> *Examples: "It's a good source of ideas. Before you despair over a task because you have no idea what to write, it's better to get different ideas. That often makes the work easier" (P.73); "AI can provide ideas that you or your group wouldn't have thought of so quickly" (P.47); "New perspectives and ideas or suggestions to broaden your horizons" (P.42).*

3.1.3 Improvement

Another frequently cited benefit was ChatGPT's role in enhancing the quality of work (19%). Teenagers mentioned its usefulness in refining text formulations, checking grammar and spelling errors, and clarifying difficult concepts. In this way, ChatGPT was not merely seen as a source of information but as a tool for enhancing the accuracy and quality of their work. Some respondents appreciated how the tool could improve their output by providing concrete examples or guiding them towards clearer, more professional expressions. One teenager even compared it to a substitute teacher, when the real one didn't have time.

> *Examples: "It can be quite useful because it can also improve your own skills" (P.40); "It's a good way to formulate your thoughts clearly and explain your own opinion in a way that's understandable for others, especially if you struggle with that" (P.5); "It's like a substitute for a teacher when they don't have time for you" (P.76).*

3.1.4 Immediacy (time and effort saving)

The immediacy with which ChatGPT delivers results was a prominent theme, with many students valuing the time and effort it saves (28%). Teenagers highlighted that ChatGPT allowed them to complete tasks more quickly and efficiently, reducing the burden of long research or writing processes. This benefit was particularly important for students facing time constraints or multiple assignments, as they felt that ChatGPT could help them manage their workload more effectively, giving them more free time for other activities.

> *Examples: "I think ChatGPT is a very interesting tool. With ChatGPT, I was able to inform myself much faster and more precisely about topics (e.g., for a presentation) <...> I think ChatGPT has the advantage of speeding up and easing many tasks." (P.119); "You can finish tasks faster and have more time for things you enjoy "(P.127); "One advantage is efficiency – you can complete tasks faster. Also, some school tasks are not really productive/useless, and you could use your time for better things" (P.94).*

3.2 Perceived Drawbacks and Concerns

Approximately 17% of teenagers reported perceiving no benefits from using ChatGPT for their homework or school projects. Several essays presented a combination of perceived advantages and concerns, illustrating a nuanced view of the tool's potential in education. The teenagers expressed the following drawbacks and concerns:

3.2.1 Lack of learning

This theme emerged as the most prominent concern in the teenagers' essays. Many responses highlighted that reliance on ChatGPT could hinder their personal learning. These students voiced worries about over-dependence on AI, fearing it may compromise their critical thinking skills and independent engagement with the material. Several teenagers articulated their fears that frequent use of ChatGPT could diminish their ability to process information and learn autonomously.

> *Examples: "I see little advantage because there is no learning process if ChatGPT does everything for you" (P.2); "It's also okay to let ChatGPT do the homework sometimes, but the use should be limited and not become a habit. Overall, ChatGPT can take away a lot of stress <…> but you shouldn't let an AI do everything. It can be useful, but it can also lead to losing important skills, and you won't be well-prepared for school tasks, etc." (P.125); "For homework, I find it very concerning because it can quickly lead to poorer performance in school since you don't repeat the content through doing homework"(P.138).*

3.2.2 Concerns about quality and accuracy

Other notable concerns centred around the quality and accuracy of the information provided by ChatGPT. A few teenagers pointed out that the AI sometimes generated incorrect information or responses that were too general to be useful for specific tasks. These concerns suggest that ChatGPT may not necessarily enhance the quality of students' work and could even contribute to poorer academic performance.

> *Examples: "I think it can lead to mistakes and that you don't learn anything from it" (P.37); "If you're doing a project where research is allowed, you could use it for summarized and concise content, but you must verify it" (P.9).*

3.2.3 Academic honesty

Ethical concerns regarding academic honesty and the potential misuse of ChatGPT to avoid learning or to cheat were also mentioned. This tension highlights the conflict between the convenience offered by the tool and the risk of undermining the educational process.

> *Examples: "you're not using your own knowledge" (P.43).*

4. Discussion

4.1 The 4I Model of Benefits

The findings indicate that most teenagers (83%) recognize practical and creative benefits of using ChatGPT for academic tasks. Analysis identified four key themes - Information, Inspiration, Improvement, and Immediacy - leading to the development of the 4I Model of Benefits (Figure 1). This model categorizes the perceived benefits of ChatGPT into four primary domains, offering an organized way to understand teenagers' views on its academic utility. It highlights that ChatGPT's attractiveness encompasses these areas: the extensive information it offers, its capacity to inspire creativity, its role in improving the quality of work, and the efficiency it brings in terms of time and effort saved.

Information ***A tool for providing quick access to information*** Provides quick access to vast resources and data, aiding in understanding and research efficiency	**Inspiration** ***A tool for creative ideas and new perspectives*** Sparks creativity by offering fresh ideas and different approaches to tasks
Improvement ***A tool for enhancing the quality of work*** Helps improve the quality of work through better text formulations, corrections, and clarifications of complex concepts	**Immediacy** ***A tool saving time and effort*** *Reduces time and effort, allowing students to complete tasks more efficiently*

Figure 1: The 4I Model of Benefits

Four types of benefits:

Information: Many teenagers highlighted that ChatGPT provides quick access to vast amounts of information, which helps them understand topics better and saves research time.

Inspiration: Students indicated that ChatGPT serves as a creative catalyst, offering ideas and perspectives that enhance their work. This observation is consistent with findings from previous studies, such as those by Sok and Heng (2023), which emphasize the tool's capacity to stimulate creativity in academic settings.

Improvement: ChatGPT is regarded by teenagers as a valuable tool for enhancing work quality, especially in areas such as text formulation, error checking, and concept clarification. Our findings align with existing research, which underscores its potential to improve writing quality and clarity. Studies indicate for instance that ChatGPT can assist in generating coherent text, identifying grammatical errors, and clarifying complex concepts in academic contexts (Grassini 2023; Bašić et al. 2023; Bem Machado et al. 2024; Mahapatra 2024).

Immediacy: Many teenagers appreciated ChatGPT for its speed and efficiency, highlighting how it reduces the effort required to complete assignments.

These insights align with existing research on AI benefits in education (Trust et al. 2023; Grassini 2023; Kasneci et al. 2023). For instance, Trust et al. (2023) identified benefits in personalized learning and creative support that align closely with the "Improvement" and "Inspiration" themes. Similarly, Sok and Heng (2023) found ChatGPT also advantageous for assessment creation, personal tutoring, and brainstorming. This aligns well with our findings, though teenagers did not directly mention assessment-related benefits, which may be more relevant for advanced or older students.

4.2 Integration with the Technology Acceptance Model (TAM)

The TAM model suggests that "perceived usefulness" and "perceived ease of use" are key factors driving technology acceptance, both shaped by external variables (Davis 1989). Figure 2 illustrates an extended TAM framework that integrates the key benefits of ChatGPT as perceived by teenagers:

- **Information, Inspiration, and Improvement** aspects directly contribute to the perceived usefulness of ChatGPT. These benefits support students in performing academic tasks more effectively and improving the quality of their work.
- **Immediacy** aligns with perceived ease of use, as ChatGPT's time-saving features reduce the effort required to complete assignments, thus increasing its appeal to students.

Together, these benefits of the 4I Model positively influence teenagers' overall attitudes toward ChatGPT, promoting its adoption for academic purposes despite potential concerns.

Our findings suggest that teenagers perceive multiple benefits to using ChatGPT for academic tasks, which are effectively captured within the 4I Model. This model adds to the existing literature by offering a structured way to categorize and understand the different types of benefits. By linking these perceived benefits to the established TAM framework, we contribute to ongoing discussions about the integration of AI in education.

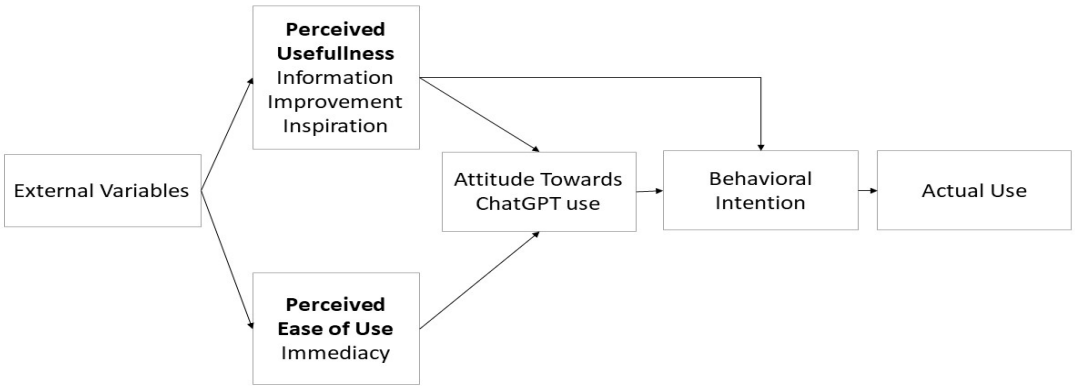


Figure 2: Integration of the 4I Model of Benefits with Technology Acceptance Model

4.3 Perceived Concerns and Implications

While 83% of teenagers in this study identified benefits to using ChatGPT for academic work, approximately 17% saw no clear advantages. The most common concern among these students was that ChatGPT might inhibit their learning. Besides potential *lack of learning*, additional concerns included issues of *inaccuracy* and *academic integrity*. These perceptions align with the broader academic discourse, which highlights both opportunities and challenges posed by generative AI in education (Kasneci et al. 2023; Trust et al. 2023; Tlili et al. 2023; Sok and Heng 2023).

While some teenagers worried that over-dependence on AI might compromise their critical thinking, one student compared ChatGPT to a “substitute teacher”. Similarly, Chan and Tsi (2023) found that although some students and teachers anticipate a larger role for AI in teaching, most believe critical thinking, creativity, and emotional insight remain uniquely human qualities that AI cannot replicate.

Beyond integrity, our findings also emphasize the importance of developing students' critical AI literacy. As AI tools like ChatGPT become more integrated into learning, students must develop the ability to question how such technologies generate information and assess the biases that may be embedded in their outputs. For example, awareness of the limitations in ChatGPT's training data - such as cultural biases or stereotypes - can help students approach AI-generated information with a critical perspective.

Overall, these insights suggest that while ChatGPT presents considerable benefits, its integration into education requires a balanced approach that proactively addresses concerns regarding potential lack of learning, academic integrity, and the need for AI literacy.

5. Limitations and Further Research

This study offers valuable insights into teenagers' perceptions of using ChatGPT for homework and school projects; however, several limitations should be acknowledged.

The sample of 141 teenagers, while geographically diverse, consists solely of German grammar school students, which may limit the generalizability of the findings to other educational systems or countries with different cultural attitudes toward AI.

This study relies on qualitative, self-reported data from written essays, which may introduce response biases, such as social desirability or the tendency to provide "correct" answers. Variations in response length and depth could also affect the thematic analysis, with some students offering more detailed insights than others.

The study's focus is limited to perceived benefits, viewed from a user-experience perspective, without direct assessment of ChatGPT's influence on learning outcomes, critical thinking, or academic integrity. It reflects reported advantages, but does not account for possible long-term effects that prolonged AI use might have on educational development.

This study uses the TAM framework to examine factors influencing ChatGPT's adoption. Future research could explore alternative models like the Value-Based Adoption Model (VAM) (Kim et al. 2007), which considers both functionality and hedonic benefits. VAM has been used to assess various technologies, including mobile internet (Kim et al. 2007) and paid learning apps (Wang et al. 2019). Al-Abdullatif and Alsubaie (2024) applied an extended VAM framework with AI literacy to evaluate ChatGPT's acceptance in higher education, suggesting a promising approach for studying AI adoption in education.

Future studies could benefit from longitudinal approaches to examine how teenagers' perceptions and usage patterns of ChatGPT evolve over time, particularly as AI technology continues to develop and integrate into educational systems. This approach could provide deeper insights into the long-term educational impact of AI generators like ChatGPT on teenagers' learning, critical thinking, and academic habits.

6. Conclusion

This study introduces the **4I Model of Benefits**, offering an original perspective on how teenagers perceive the benefits of ChatGPT in academic contexts. It identifies **four key benefits**: *Information, Inspiration, Improvement*, and *Immediacy*. By positioning this model within the TAM framework, we gain a deeper understanding of the external factors that drive the acceptance of AI text generators among students.

The 4I Model of Benefits contributes to the growing discourse on the role of AI in education and provides a useful tool for future research on integrating generative AI into educational settings. While our findings clearly outline the advantages of using ChatGPT, they also emphasize the need for balanced usage. Over-reliance on AI tools may hinder the development of essential learning skills and critical thinking. Future research should explore strategies for effectively integrating these technologies into educational practices, aiming to maximize their benefits while minimizing potential drawbacks.

References

Abramski, Katherine; Citraro, Salvatore; Lombardi, Luigi; Rossetti, Giulio; Stella, Massimo (2023): Cognitive Network Science Reveals Bias in GPT-3, GPT-3.5 Turbo, and GPT-4 Mirroring Math Anxiety in High-School Students. In: *Big Data and Cognitive Computing* 7 (3), p. 124.

Al-Abdullatif, Ahlam Mohammed; Alsubaie, Merfat Ayesh (2024): ChatGPT in Learning: Assessing Students' Use Intentions through the Lens of Perceived Value and the Influence of AI Literacy. In: *Behavioral sciences* 14 (9).

Azaria, Amos; Azoulay, Rina; Reches, Shulamit (2024): ChatGPT is a Remarkable Tool - For Experts. In: *Data Intelligence* 6 (1), p. 240–296.

Bagozzi, Richard (2007): The Legacy of the Technology Acceptance Model and a Proposal for a Paradigm Shift. In: *Journal of the Association for Information Systems* 8 (4), p. 244–254.

Bašić, Željana; Banovac, Ana; Kružić, Ivana; Jerković, Ivan (2023): ChatGPT-3.5 as writing assistance in students' essays. In: *Humanities and Social Sciences Communication* 10 (1).

Bem Machado, Andreia de; Pesqueira, Antonio; Sousa, Maria José (2024): ChatGPT and the Future of Education: A Critical Review. In: Andreia de Bem Machado, Maria José Sousa, Francesca Dal Mas, Silvana Secinaro and Davide Calandra (Hg.): Digital Transformation in Higher Education Institutions. Cham: Springer Nature Switzerland, p. 155–176.

Bernardone, Leonard (2023): ChatGPT rattles Australian education sector. Burgeoning AI tool banned in public schools. In: *InformationAge: ICT News* (Feb 02) https://ia.acs.org.au/article/2023/chatgpt-rattles-australian-education-sector.html Last accessed 24.10.2024.

Bhattacharyya, Mehul; Miller, Valerie M.; Bhattacharyya, Debjani; Miller, Larry E. (2023): High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. In: *Cureus* 15 (5).

Chan, Cecilia Ka Yuk; Tsi, Louisa H. Y. (2023): The AI Revolution in Education: Will AI Replace or Assist Teachers in Higher Education? ArXiv preprint arXiv:2305.01185.

Choi, Jonathan H.; Hickman, Kristin E.; Monahan, Amy; Schwarcz, Daniel B. (2022): ChatGPT Goes to Law School. In: *Journal of Legal Education 71 (3)*, p. 387- 400.

Cotton, Debby R. E.; Cotton, Peter A.; Shipway, J. Reuben (2023): Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. In: *Innovations in Education and Teaching International*, p. 1-12.

Davis, Fred D. (1989): Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. In: *MIS Quarterly* 13 (3), p. 319.

DuBoulay, Benedict; Mitrovic, Antonija; Yacef, Kalina (Hg.) (2023): Handbook of artificial intelligence in education. Cheltenham, Northampton: Edward Elgar Publishing.

Farrokhnia, Mohammadreza; Banihashem, Seyyed Kazem; Noroozi, Omid; Wals, Arjen (2024): A SWOT analysis of ChatGPT: Implications for educational practice and research. In: *Innovations in Education and Teaching International* 61 (3), p. 460–474.

Grassini, Simone (2023): Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. In: *Education Sciences* 13 (7), p. 692.

Hagendorff, Thilo (2020): The Ethics of AI Ethics: An Evaluation of Guidelines. In: *Minds & Machines* 30 (1), p. 99–120.

Holmes, Wayne; Bialik, Maya; Fadel, Charles (2019): Artificial intelligence in education. Promises and implications for teaching and learning. Boston, MA: Center for Curriculum Redesign. https://permalink.obvsg.at/AC17102554. Last accessed 24.10.2024.

Kasneci, Enkelejda; Sessler, Kathrin; Küchemann, Stefan; Bannert, Maria; Dementieva, Daryna; Fischer, Frank et al. (2023): ChatGPT for good? On opportunities and challenges of large language models for education. In: *Learning and Individual Differences* 103, p. 102274.

Kim, Hee-Woong; Chan, Hock Chuan; Gupta, Sumeet (2007): Value-based Adoption of Mobile Internet: An empirical investigation. In: *Decision Support Systems* 43 (1), p. 111–126.

Kung, Tiffany H.; Cheatham, Morgan; Medenilla, Arielle; Sillos, Czarina; Leon, Lorie de; Elepaño, Camille et al. (2023): Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. In: *PLOS digital health* 2 (2).

Mahapatra, Santosh (2024): Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study. In: *Smart Learning Environments* 11 (1).

Meyer, Jesse G.; Urbanowicz, Ryan J.; Martin, Patrick C. N.; O'Connor, Karen; Li, Ruowang; Peng, Pei-Chen et al. (2023): ChatGPT and large language models in academia: opportunities and challenges. In: *BioData mining* 16 (1), p. 20.

Miao, Fengchun; Holmes, Wane; Huang, Ronghuai; Zhang, Hui (2021): AI and education: guidance for policy-makers: UNESCO. https://permalink.obvsg.at/. Last accessed 24.10.2024.

Nguyen, Andy; Ngo, Ha Ngan; Hong, Yvonne; Dang, Belle; Nguyen, Bich-Phuong Thi (2023): Ethical principles for artificial intelligence in education. In: *Education and information technologies* 28 (4), p. 4221–4241.

Ogugua, Divine; Yoon, Seong No; Lee, DonHee (2023): Academic Integrity in a Digital Era: Should the Use of ChatGPT Be Banned in Schools? In: *Global Business and Finance Review* 28 (7), p. 1–10.

OpenAI.com (2024): Terms of use. https://openai.com/policies/row-terms-of-use/. Last accessed 24.10.2024.

Pavlik, John V. (2023): Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. In: *Journalism & Mass Communication Educator* 78 (1), p. 84–93.

Pegoraro, Alessandro; Kumari, Kavita; Fereidooni, Hossein; Sadeghi, Ahmad-Reza (2023): To ChatGPT, or not to ChatGPT: That is the question! ArXiv preprint arXiv:2304.01487.

Rahman, Md. Mostafizer; Watanobe, Yutaka (2023): ChatGPT for Education and Research: Opportunities, Threats, and Strategies. In: *Applied Sciences* 13 (9), p. 5783.

Ravšelj, Dejan; Aristovnik, Aleksander; Keržič, Damijana; et al. (2024): Higher Education Students' Early Perceptions of ChatGPT: Global Survey Data. Manuscript in preparation for submission to PLoS ONE.

Rozado, David (2023): The Political Biases of ChatGPT. In: *Social Sciences* 12 (3), p. 148.

Rudolph, Jürgen; Crawford, Joseph; Sam, Choon-Yin; Tan, Shannon (Hg.) (2024): The Palgrave Handbook of Crisis Leadership in Higher Education. Cham: Springer Nature Switzerland.

Sallam, Malik (2023): ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. In: *Healthcare* 11 (6).

Sok, Sarin; Heng, Kimkong (2023): ChatGPT for education and research: A review of benefits and risks. Available at SSRN: https://ssrn.com/abstract=4378735. Last accessed 24.10.2024.

Sullivan, Miriam; Kelly, Andrew; McLaughlan, Paul (2023): ChatGPT in higher education: Considerations for academic integrity and student learning. In: *Journal of Applied Learning and Teaching,* 6 (1), p.1-10.

Tlili, Ahmed; Shehata, Boulus; Adarkwah, Michael Agyemang; Bozkurt, Aras; Hickey, Daniel T.; Huang, Ronghuai; Agyemang, Brighter (2023): What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. In: *Smart Learning Environments* 10 (1).

Trust, Torrey; Whalen, Jeromie; Mouza, Chrystalla (2023): ChatGPT: Challenges, Opportunities, and Implications for Teacher Education. In: *Contemporary Issues in Technology and Teacher Education* (23/1), p. 1–23.

Venkatesh, Viswanath; Morris, Michael; Davis, Fred (2003): User Acceptance of Information Technology: Toward a Unified View. In: *MIS Quarterly* 27 (3), p. 425.

Venkatesh, Viswanath; Davis, Fred D. (2000): A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. In: *Management Science* 46 (2), p. 186–204.

Wang, Yu-Yin; Wang, Yi-Shun; Lin, Hsin-Hui; Tsai, Tung-Han (2019): Developing and validating a model for assessing paid mobile learning app success. In: *Interactive Learning Environments* 27 (4), p. 458–477.

Wen, Jun; Wang, Wei (2023): The future of ChatGPT in academic research and publishing: A commentary for clinical and translational medicine. In: *Clinical and translational medicine* 13 (3), 1207.

Young, Jeffrey R. (2023): AI Tools Like ChatGPT May Reshape Teaching Materials - And Possibly Substitute Teach. In: *EdSurge: Teaching and Learning* (January 19). https://www.edsurge.com/news/2023-01-19-ai-tools-like-chatgpt-may-reshape-teaching-materials-and-possibly-substitute-teach. Last accessed 24.10.2024.

Yu, Hao (2023): Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. In: *Frontiers in psychology* 14, 1181712.

Adopting Artificial Intelligence in Organisations: A Closer Look

Massimo Albanese
Department of Economics and Business Studies, University of Genoa, Italy

Massimo.Albanese@unige.it

Abstract: Artificial intelligence (AI) is increasingly being adopted in different types of organisations and is attracting the attention of various actors. In this context, the analysis aims to provide an overview of the most relevant aspects of the adoption of AI technology solutions in organisations. To this end, the analysis adopted the archival research method and carried out a study of recent documents/analyses proposed by leading consulting firms. These players accumulate considerable application knowledge and are, therefore, special sources. To provide an in-depth and up-to-date picture, the analysis considered a range of archival information (such as reports, articles, insights, technical notes, etc.) published by around 15 organisations in the consulting sector since the beginning of 2023. The study highlights the main opportunities/risks associated with the introduction of AI. The findings cover various aspects, including AI investments, classification of organisations adopting AI, main challenges in AI initiatives, most common areas of AI use, and changes in work. The results may have some limitations due to bias in the identification of documents, as they were derived from direct searches and queries on the search engines of the consulting firms' websites. The audience and implications are broad, and this is a value of the work. AI novices studying the adoption and implementation of AI have a picture of the relevant dimensions on which further in-depth studies can be developed. In addition, entrepreneurs, managers, and policymakers will have an overview of the main threats/opportunities of AI and elements to support decision-making.

Keywords: Artificial intelligence, Adoption, Organisation, Archival research

1. Introduction

In the past, new technological solutions have led to breakthroughs and waves of progress. One might cite the internal combustion engine as an example of a technology that contributed to the Industrial Revolution at the end of the 19th century. Although AI is not new - Turing first conceptualised it in 1950 and McCarthy coined the term artificial intelligence in 1955 - we may be on the cusp of a new era that will be shaped by the pervasive impact of artificial intelligence. This can be defined, in the most basic terms, as technological solutions that emulate or imitate human intelligence and behaviour. In other words, machines that perform cognitive functions normally associated with the human mind.

There has been a notable increase in the adoption of AI in recent years. Major companies are investing heavily in AI, and venture capitalists are excited to invest in new disruptive startups. Indeed, it is estimated that investments in AI will approach $160-200 billion globally by 2025 (Goldman Sachs, 2023a). A variety of stakeholders, including practitioners, managers, politicians, and trade unions, are focusing on AI technologies, as evidenced by the growing number of mentions in earnings call transcripts. In the first quarter of 2022, there were over 5,000 mentions, while in the third quarter of 2023, this number had risen to approximately 37,000 (Daugherty et al., 2024). It is therefore unsurprising that some authors view AI as a general-purpose technology (like the steam engine, electricity, etc.), which is developing rapidly and is already pervasive, capable of driving complementary innovations and creating many opportunities.

In this sense, generative AI[1] (GenAI) represents one of the technologies that has the potential to bring about significant advancements. Indeed, some models estimate that the adoption of responsible GenAI could potentially unlock an additional economic value (compared to the expected baseline growth) of $10.3 trillion by 2038 (Shook and Daugherty, 2024). Similarly, a considerable number of managers posit that GenAI has the potential to augment companies' market share in the short term, with growth possibilities reaching 10% or more. Additionally, projections indicate that there will be enhancements in productivity, creativity, and human potential.

The adoption of AI poses several challenges for organisations, which need to consider many issues. Does my organisation need AI? How will AI increase efficiency? Will AI alone deliver certain improvements? How will AI change my organisation? Will AI be reliable? There are, of course, many other questions that could be asked. However, it seems that a common denominator is the way in which organisations can drive and control AI.

[1] GenAI has the potential to create new content and accelerate research and development cycles, as it is based on AI techniques that learn from artefacts and data and generate innovative new creations that are similar but not identical to the original.

With this in mind, this paper seeks to provide an overview of the most salient aspects of AI in organisations by analysing recent documents proposed by leading consulting firms (Accenture, BCG, EY, Gartner, McKinsey, etc.). As these actors possess considerable application knowledge, they represent a valuable source of information on the subject. Consequently, the paper provides valuable insights into the key characteristics of the operational and strategic levels, as well as the potential opportunities and risks associated with the adoption of AI. Accordingly, the analysis covers a variety of aspects, including areas of AI use, the impact of AI initiatives, the classification of adopting organisations and types of projects.

The audience and implications are broad, which represents a valuable aspect of the work. It provides entrepreneurs, executives, and policymakers with an overview of AI and a set of elements to support their decisions on AI adoption. At the same time, researchers who are less familiar with AI may find this a useful starting point for further in-depth studies.

The paper is structured as follows. The second section provides an overview of the knowledge base concerning the adoption of AI in organisations. The third section presents the main transformations and impacts that AI solutions bring about in organisations. In the concluding section, the findings are presented in the form of recommendations and suggestions for supporting the adoption of AI in organisations.

2. AI in Organisations: Some Basics

The lack of a universally accepted definition of AI has the potential to confuse those without expertise in the field. The term itself is broad, and therefore defining it may prove challenging. Firstly, it is useful to acknowledge that AI is not merely a technology, but also a discipline. In particular, it is a branch of computer science that aims to develop machines with the capacity to perform tasks through problem-solving, learning and understanding natural language. This implies the processes of perception, reasoning, and decision-making, and the capacity to act. As a result, AI draws on a range of subfields and techniques, including machine learning, neural networks and deep learning, which are essential for the realisation of AI systems.

As a technology, "artificial intelligence refers to systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or internet of things applications)" (European Commission, 2018, p.1). In what follows, the analysis views AI as a technology and uses the term AI to refer to all AI-based technologies. It is inaccurate to view AI as a single technology; rather, it should be regarded as a constellation of technologies that collectively enable designed solutions to sense, comprehend, learn, and act, resembling that of human capabilities.

There are numerous applications of AI, which can be classified in a variety of ways. In summary, applications can be classified as either "weak AI" (artificial narrow intelligence), which is designed and trained to perform specific tasks (e.g., voice recognition), or "strong/full AI" (artificial general and superintelligence), which is characterised by cognitive abilities equal to or greater than those of humans. The latter category gives rise to concern, as some experts claim that there is a potential for AI to surpass human intelligence. Considering this calls for new regulations are becoming increasingly prevalent, and new concepts are emerging, such as responsible AI[2].

However, when considering AI as a technology, the literature can be broadly divided into two distinct streams: those dealing with the design and development of AI solutions and those dealing with the adoption and use. As the latter stream is the focus of this paper, the current analysis refers to the literature on the diffusion of innovation theory, the technology acceptance model, and the technology-organisation-environment framework.

A considerable body of economic literature has focused on the characteristics of firms that facilitate the adoption of new technologies. These include factors such as organisational size, readiness, workforce skills, adoption of complementary technologies, sector of activity, and competitive pressure. However, technology features also exert an influence on the adoption. These include perceived usefulness, which is the perception of the extent to which the use of a particular technology improves performance, and perceived ease of use, which refers to the extent to which individuals believe that using a particular technology is effortless (Davis,

[2] This refers to the process of developing and using AI solutions in accordance with ethical values and organisational objectives

1989)[3]. These factors elicit cognitive responses that contribute to the formation of a positive attitude and behavioural intention to adopt a particular technology (Davis, 1989).

In summary, the more a technology is perceived as straightforward to utilise, the more probable it is that the user will perceive it as beneficial, and the more likely it is that the technology will be adopted. This is also the case with AI, as indicated by Kelly et al. (2023, p.1), who reviewed the literature on the acceptance of AI solutions and claimed that "perceived usefulness, performance expectancy, attitudes, trust, and effort expectancy significantly and positively predicted behavioural intention, willingness, and use behaviour of AI across multiple industries".

Other factors may hinder the adoption and utilisation of technologies, such as infrastructure and regulatory gaps or lack of data. When viewed as systems, AI solutions typically comprise four key components. These are energy, computation, data and models. This allows for further considerations to come to light. Energy is a fundamental ingredient in the training and operation of AI systems. There is an increasing demand for computation, yet the supply is sometimes inadequate. There is also an increasing demand for data, as data is essential for any form of computation, particularly for AI. Large and small models (such as GPT-4, Bard, Falcon, etc.) are evolving and it is useful to link them to more resources to increase their potential. It is regrettable that the factors affecting the adoption and utilisation of AI are numerous, making it impractical to present them in this paragraph. Accordingly, this overview is provided as an initial introduction to the subject matter.

2.1 AI as a General-Purpose Technology: Laggards and Transformers

Historically, new technological solutions that have resulted in significant breakthroughs and waves of progress have typically followed an S-shaped pattern. In the view of some authors, AI will play a similar role to that of steam in the first industrial revolution, thereby characterising the fourth[4]. Indeed, AI exhibits numerous similarities with preceding disruptive general-purpose technologies. Firstly, AI has demonstrated a remarkable capacity for rapid improvement. The growth of AI has been exponential, as is its ability to process data. Furthermore, as many surveys have demonstrated, AI is pervasive. Finally, AI can foster complementary innovation by allowing users to recombine and exploit existing technologies and knowledge to create new solutions.

Although AI is not a novel concept, it has only become mainstream in the last few decades as large technology companies have started to integrate AI into various products[5]. That said, the recent evolution of AI can be divided into three phases (Shook and Daugherty, 2024). The first phase, the diagnostic phase, is characterised by the introduction of machine learning. The second phase, the predictive phase, is focused on forecasting. The third phase, the generative phase, is distinguished by AI's ability to not only accurately predict but also to create new content and offer customised suggestions.

However, the adoption of AI has not been uniform. A considerable number of organisations have invested in numerous standalone use cases that have frequently demonstrated the value of AI, yet only a small proportion of initiatives can be considered to have reached a fully developed stage. Consequently, current AI projects exhibit a range of levels of maturity. On this basis, it is possible to categorise organisations as follows:

- Laggards, those who have not yet taken the necessary steps to adopt AI,
- Enthusiasts, organisations equipped with the basic technology to experiment with AI,
- Organisations on the journey, those with the technology and expertise to initiate AI projects,

[3]This factor is similar to the complexity factor that characterises the innovation diffusion literature, which refers to the level of difficulty people perceive in understanding and using new technologies.

[4]The first industrial revolution commenced at the conclusion of the 18th century, propelled by the advent of coal and the subsequent transformation of our economy from an agrarian to an industrial base. The second industrial revolution began at the end of the 19th century. It was driven by new sources of energy, namely electricity, gas and oil, and the invention of the combustion engine. The third industrial revolution started in the second half of the 20th century. It was characterised by the emergence of nuclear energy and the development of electronics and telecommunications. The fourth industrial revolution commenced at the beginning of the third millennium. It is characterised by the acceleration of digital technologies and the internet (including the cloud technology, Internet of Things and AI), as well as a shift towards renewable energy.

[5]Apple is a noteworthy example, having introduced Siri, a voice assistant that uses AI to understand language, in 2011.

- Apprentices, those who have the technology and need to improve their skills to make AI projects fully operational, and
- Cutting-edge adopters, organisations that have the technology and expertise to successfully implement and deploy AI.

In particular, an analysis of organisations that have adopted AI reveals four distinct categories based on the number of AI applications that have been fully deployed and the high-level outcomes achieved (Mittal et al., 2022). These categories are as follows: a) Starters (limited applications deployed and low outcomes achieved), b) Underachievers (high deployment and low outcomes achieved), c) Path seekers (limited applications deployed and high outcomes achieved, which implies success in fewer initiatives), and d) Transformers (high deployment and high outcomes achieved). Numerous transformers are demonstrating superior performance in a multitude of domains when compared to their competitors. Furthermore, the discrepancy between the leading transformers and the laggards is increasing as the capabilities in which the former have invested continue to enhance. Consequently, there is a possibility that the laggards, particularly small and medium-sized enterprises, may be unable to maintain pace.

2.2 Main Challenges and Applications

The adoption of AI solutions presents several significant challenges, and organisations must therefore be prepared to address specific pivotal aspects:

- Supply chain. With the exponential growth in demand for data, organisations need actionable data and internal collaboration, given that the mapping, monitoring and management of supply chains can be complex,
- Strategy. AI projects need advocacy and accountability from top management. Without strategic guidance, AI solutions can be misaligned, and AI initiatives can stall,
- People. As AI requires skilled people, there may be a mismatch between the skills required and the current workforce. This can create a challenging situation where there is uncertainty about the skills required, the availability of talent is limited, and upskilling and training programmes are required, and
- IT. AI requires the right infrastructure and the right environment. Inadequate resource allocation and technology architecture can therefore hamper AI initiatives. A typical case is where IT departments have small teams with insufficient resources.

The most prevalent use cases for AI in organisations are the automation of IT tasks, cybersecurity, fraud detection, business processes and workflow automation. These are the areas in which organisations are currently utilising or planning to utilise AI in the future. While some tasks will not be affected by AI, much of the work will be assisted and automated. As a result, humans working with AI co-pilots will increase the tasks performed and the results achieved. In this regard, it is beneficial to differentiate between two principal applications of AI in organisations, namely automation and augmentation.

The concept of automation implies that AI performs a human task (in this case, humans delegate tasks to AI with little or no involvement), whereas the concept of augmentation implies that humans work closely with (and are assisted by) AI to perform tasks. In this case, humans and AI are not adversaries, and their relationship is no longer seen as dichotomous. On the contrary, they interact constantly.

Tensions may occasionally emerge when organisations are required to select between automation and augmentation. However, some authors argued that organisations should prioritise augmentation, as it can deliver superior performance. Other authors suggested that organisations should adopt an approach that includes both automation and augmentation, as this allows for complementarities that are beneficial to the business (augmentation can increase the level of automation). In addition, some authors have argued that automation and augmentation cannot be neatly separated.

AI can be used to achieve many benefits (such as increasing efficiency, customising or improving existing products/services, predicting demand, detecting fraud, managing risks, and more). Consequently, there exists a diverse array of potential application areas. At present, some organisations prioritise productivity objectives and utilise AI primarily to automate/augment front and back-office operations. Some exclude the use of AI in front of customers and only employ it internally, whereas others pursue the full use of AI by adopting it across the entire organisation. To categorise AI applications, Gartner proposed three types of use cases (Wiles, 2023):

- Quick wins. Such initiatives tend to be task-specific, easy to buy, easy to launch, and easy to pilot. The implementation of a productivity assistant, such as Microsoft 365 Copilot, is a typical example. These solutions can help organisations gain a competitive advantage. Their time-to-value is short (typically within a year), and the improvements can be a source of differentiation, but they diminish over time,
- Differentiating use cases. These initiatives offer a more defensible advantage than in the previous case, as they allow organisations to improve processes by using data in unique ways. They may involve a process redesign, skill upgrading, and risk management, making costs and risks more unpredictable. However, the direct and indirect benefits generally outweigh the costs, and the time-to-value is medium (generally between one and two years), and
- Transformative initiatives. These projects have the potential to deliver greater benefits but are difficult to value in financial terms. Although ongoing developments in AI are reducing the cost of deployment, executives and boards tolerate more risk and a longer time to value (also more than two years) in exchange for a first-mover advantage.

3. Impacts, Risks, and Transformations Across Organisations

The deployment of AI projects can necessitate considerable resources. However, it is probably not the size of the investment that is the primary obstacle, but an investment with less transparent benefits and costs, such as the cost of people, the cost of the infrastructure required, or the cost of using specific AI solutions, and therefore the uncertainty regarding the return on investment. In light of this, the following findings from a recent global IDC study (Jyoti and Schubmehl, 2023) are worthy of note: a) 92% of AI deployments require one year or less, b) on average, organisations realise $3.5 in return for every $1 invested in AI, c) on average, organisations realise a return on investment within 14 months of deployment and d) 43% of organisations intend to reduce spending in other business areas to reallocate to AI within 24 months.

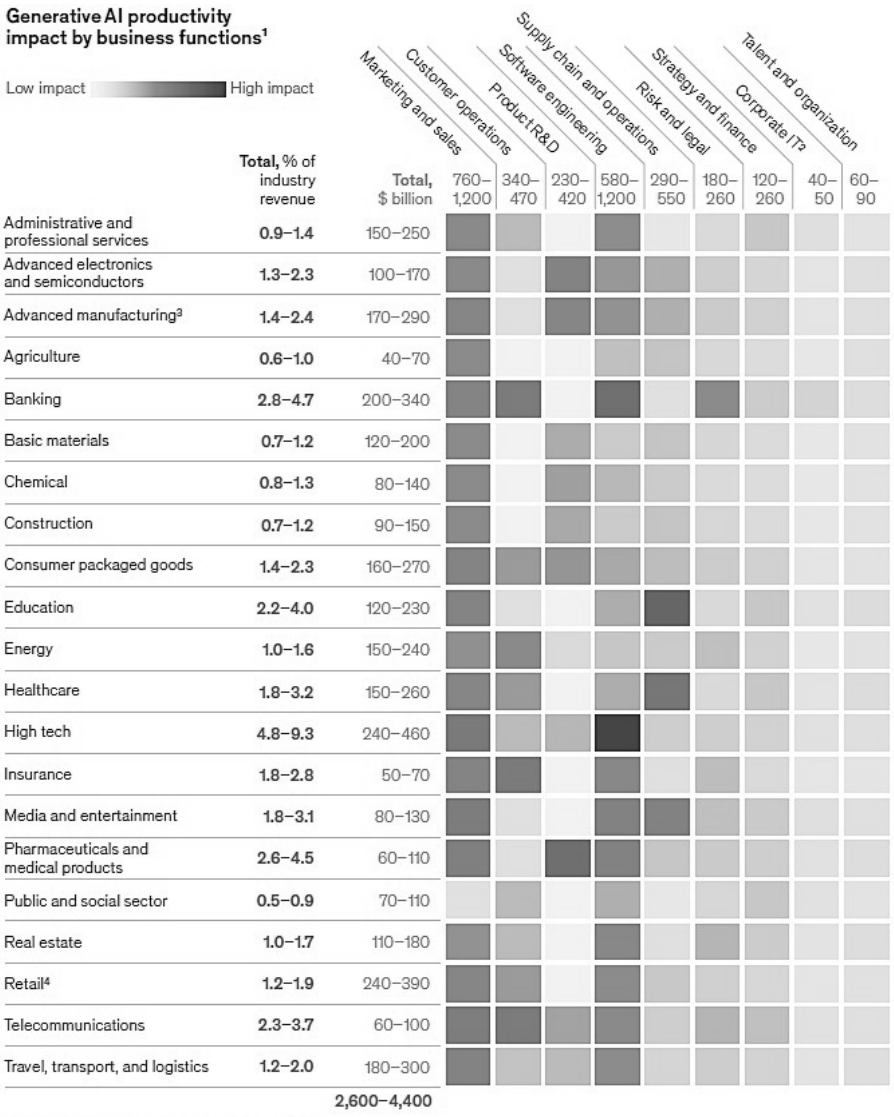
However, the impact of AI on organisations can vary, but in many cases, it can be significant. In particular, given the wide range of tasks that GenAI can perform, organisations have high expectations for this type of solution. In this regard, a McKinsey report (2024) posits that, based on 63 GenAI use cases, these initiatives have the potential to generate between $2.6 trillion and $4.4 trillion in revenue across industries (Figure 1). Furthermore, the utilisation of AI can also lead to significant cost reductions (Figure 2). For instance, Goldman Sachs (2023b) indicates that approximately 25% of the current workload in the United States and Europe could be automated. A differentiated impact of AI can be influenced by the automation/augmentation of work across sectors and countries. In this regard, Shook and Daugherty (2024) present estimates of the proportion of working hours that could be affected by automation or augmentation through GenAI, broken down by country. These estimates range from 31% in India to 47% in the UK. Estimates by Bain & Company (2023) corroborate this view by delineating roles that can be automated and augmented by GenAI (Figure 3).

Looking ahead, Statista (2023) (based on estimates from Accenture and Frontier Economics) notes that AI has the potential to increase labour productivity. In developed countries, the impact is expected to range from +11% in Spain to +37% in Sweden by 2035.

However, it would be erroneous to conclude that the advent of AI in organisations is wholly positive. Indeed, there are some risks associated with its implementation. When considering AI, inaccuracy, cybersecurity, and intellectual property infringement are the most frequently cited risks (McKinsey, 2023)[6]. This underscores the rationale behind the establishment of policies to govern the use of GenAI and the efforts to mitigate the associated risks. Some AI solutions (such as foundation models) comprise innumerable parameters and are trained on petabytes of data. Consequently, if the data used to train the model is biased, the resulting output may also be suboptimal. Additionally, models have the potential to generate misleading information, a consequence commonly referred to as “hallucinations”. Consequently, organisations must be aware of the necessity to oversee model training and assess the results, as AI solutions may require adjustment before they are used.

It is of particular importance to assess the ethical and legal implications of the outputs produced by AI. For instance, AI could potentially lead to discriminatory outcomes and inadvertently infringe the copyright of third parties. Furthermore, the use of big data and cloud computing may violate data protection regulations. It is essential to exercise caution not only with regard to the utilisation of data, but also with respect to its storage.

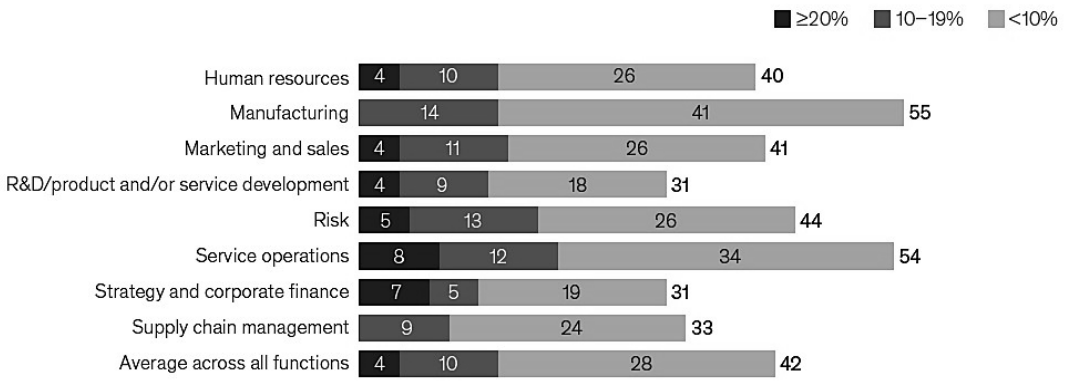
[6] The survey was conducted on a sample of 913 respondents.



Source: McKinsey (2004)

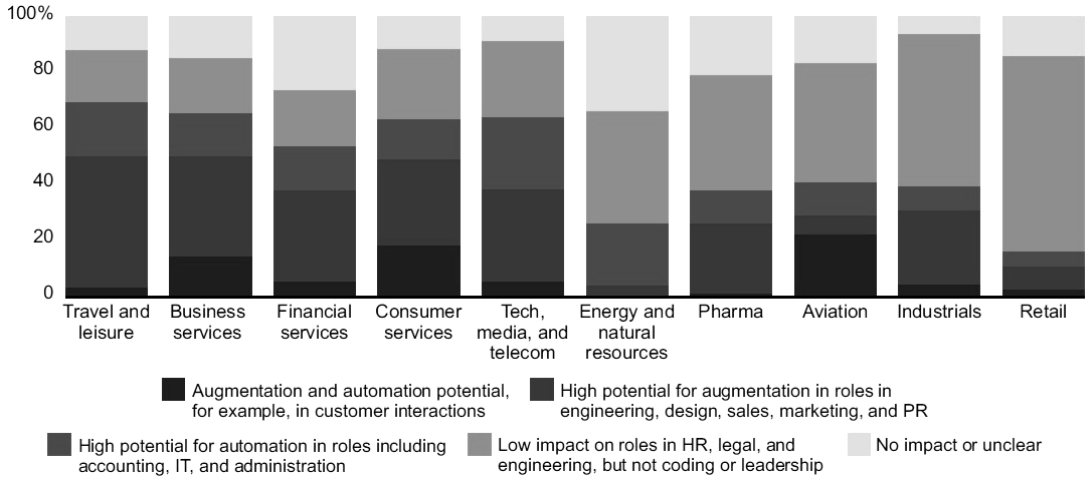
Figure 1: Impact on business functions across industries of different GenAI use cases

Notes: Figures may not sum to 100%, because of rounding. [1] Excludes implementation costs (e.g., training. Licenses); [2] Excludes software engineering; [3] Includes aerospace, defence, and auto manufacturing; [4] Includes auto retail.



Source: McKinsey (2023)

Figure 2: Cost decreases from AI adoption in 2022 (% of 1,684 respondents)



Source: Aura, Bain analysis (Bain & Company, 2023)

Figure 3: Impact of GenAI depending on automatable and augmentable roles (% of employees by function level of automation/augmentation potential)

There are also other consequences to consider. To illustrate, GenAI requires substantial electrical power. Although not all automated tasks will lead to layoffs, there are concerns about the potential unemployment caused by AI. Those tasks which are characterised by high accuracy and low context variability (such as data entry) are at risk of automation. In addition, tasks with moderate creative difficulty, context variability and accuracy are well suited to GenAI. For these reasons, 58% of workers[7] have expressed concern about job security (Shook and Daugherty, 2024).

This potential risk could be mitigated by the emergence of new AI-related occupations (data scientists, software and data engineers, software developers, etc.), as historical evidence indicates that technological innovation has consistently led to the creation of new job roles[8]. This is evident from the growing demand for technical expertise and the significant increase in job postings since 2018. Nevertheless, it is difficult to evaluate the overall impact, as a proportion of employees will undergo reskilling and new technologies will facilitate these changes.

The advent of AI (especially GenAI) will require significant changes in organisations, compelling them to modernise their technological architectures to determine the optimal organisational configuration and to ensure the responsible use of AI. AI will have a transformative impact on the way work is done. To cite one example, in the area of monitoring and control activities, organisations will be able to use more effective descriptive KPIs, as well as new and improved predictive and prescriptive metrics. The application of AI will facilitate the ongoing refinement of existing metrics, and given the complex interplay between performance drivers, AI will ensure a more comprehensive understanding of the relationship between KPIs, thereby enabling the development of new interpretations of the relationships between operational drivers, contextual variables, and results.

In summary, organisations that adopt AI are more likely to enhance their performance in terms of efficiency and effectiveness. However, if they are unable to overcome typical obstacles (such as a lack of skilled staff, difficulty in identifying relevant use cases, unavailability and/or poor quality of data, etc.), transformational outcomes (such as business model innovation) may be delayed.

4. Recommendations and Final Remarks

At present, few organisations use AI to its full potential. This raises the question of whether organisations are adequately prepared for a new digital transformation. Many reports show that there are not just technical, but also cultural, collaborative, coordination and communication obstacles to overcome.

[7]The research ("Accenture Change Workforce Survey") is based on 5,000 workers.

[8]Autor et al. (2024) estimate that around 60% of today's workers are employed in jobs that didn't exist in the 1940s.

To assess the viability of AI projects, organisations must consider several specific dimensions. These include: a) technical feasibility, which pertains to the capacity of organisations to acquire and deploy AI solutions, b) internal readiness, which concerns the receptivity and capability of organisations to utilise and harness the potential of AI, and c) external readiness, which relates to the degree of acceptance of AI among key stakeholders, including customers, suppliers, and other external parties.

There is a notable lack of preparedness, particularly when considering a disruptive solution such as GenAI. Indeed, 78% of the leaders surveyed[9] believe that their organisations are not adequately prepared to address talent issues. This sentiment is echoed by 75% of respondents in relation to governance and risk concerns. Moreover, 66% of respondents perceive deficiencies at the strategic level, while 59% indicate constraints in the technology infrastructure (Dutt et al., 2024).

Challenges also vary according to the phase of the AI project. For instance, during the initial stages of an AI project, the primary challenge is to demonstrate the business value of the initiative. Subsequently, during the scale-up phase, projects can be influenced by factors such as risk management, AI maintenance, or integration requirements. Thus, the development of an AI-enabled organisation is significantly influenced by clear leadership, discipline, and effective investment.

Given the difficulty of providing a complete guide, this analysis offers suggestions and recommendations for supporting AI adoption in organisations.

In order to ensure the effective orchestration and development of AI initiatives, it is of the utmost importance that organisations possess the necessary capabilities to mobilise resources and acquire relevant skills. This is essential to prevent the emergence of disjointed and misaligned projects and to ensure that AI is used in a beneficial and responsible manner. It is therefore recommended that organisations be prepared in terms of the following:

- Strategy. An effective strategic roadmap can facilitate the alignment of management on AI projects and enable appropriate investments that create a competitive advantage.
- Talent. The availability of the appropriate combination of qualified personnel is essential for the development of innovative applications of AI, as it enables the understanding of business models and data.
- Operating model. The development of an appropriate operating model enables cross-functional teams to function more effectively, thereby increasing the productivity of the organisation.
- Technology. A distributed technology environment that facilitates access to data, applications, and software development tools has the potential to accelerate the pace of innovation.
- Data. Given that data are typically a source of risk, it is more likely to be dangerous if organisations focus solely on the volume of data. To mitigate risks, it is prudent to invest in identifying data needs, but most importantly, data quality must be prioritised. Furthermore, ongoing data enrichment and the development of a data architecture that focuses on packaged and reusable data will facilitate access to and use of data, thereby enhancing business performance.
- Change management. Planning, resourcing and managing transformation enable organisations to adopt and scale AI solutions and optimise value capture.

It is probable that, initially, few organisations will be able to meet all the requirements. Consequently, to avoid becoming a “laggard” or remaining a “starter”, it would be advised that they adopt a two-speed AI approach, based on the logic of “start small” and “develop later”. The “start small” option may entail the implementation of modular, low-impact projects with minimal investment and tangible results in the short term. Indeed, many consulting firms often recommend starting with a limited scope to facilitate the successful implementation of AI. This involves carefully selecting use cases and trying to understand the impact and implications of AI adoption within the context of the business and the organisation. In this sense, “quick wins” projects are the typical examples of “start small” decisions. As the benefits of these projects become apparent, the initial “start small” initiative is evaluated and may evolve into more complex projects (“develop later” option) through appropriate capital allocation and strategic orchestration.

[9]The survey was conducted on a sample of 2,835 respondents, and the aforementioned percentages represent the sum of responses indicating a lack of preparation (“not prepared”), a slight level of preparation (“slightly prepared”), and a moderate level of preparation (“moderately prepared”).

These two options can be linked to two types of investment. The first type is the “no regret” investment, which focuses on a specific objective, such as productivity gains. The second type is the “strategic bets”, which aims to create new competitive advantages. The value of a two-speed AI approach does not lie in the provision of two alternatives, but rather in the benefits of their integrated design. To exemplify, it is essential to demonstrate the value of AI initiatives during the initial stages. Consequently, from a “start small” perspective, organisations can choose to establish an AI value realisation office through a management office with a limited scope, which could be considered a “no regret” investment. Subsequently, when the advantages of AI are evident and well-established, this office can evolve into a kind of control centre (“strategic bets”). This could take the form of an office with greater credibility, a broader scope, more resources, and greater autonomy and responsibility, not only for establishing specific AI projects but also for workforce reskilling plans, data infrastructure upgrades, and ecosystem strategies. Such an approach helps organisations in both incremental thinking and the development of an AI vision. Consequently, “start small” decisions should not only be regarded as initiatives to enhance efficiency and/or effectiveness. They should also be perceived as opportunities to prepare organisations for the advent of the AI era.

Effective communication and a high level of transparency are essential from the outset of AI projects. It is vital that employees are informed of the potential impact of AI on employment, including both the creation of new roles and the loss of existing ones. Furthermore, they should be made aware of the training plans in place to prepare them for future roles. However, it is not sufficient to focus solely on internal matters. For example, prioritising the relationship between AI and humans also entails monitoring other aspects, such as how customers perceive and feel about AI. In conclusion, providing individuals with the appropriate support (for example, by facilitating tailored learning pathways that meet each employee's aspirations and needs) can prove beneficial in terms of revenue, productivity, creativity, and human potential (Shook and Daugherty, 2024). Such an approach facilitates the creation of an AI-friendly culture. This is a significant element in fostering a positive attitude among employees and teams. It is therefore crucial to conduct a thorough analysis of the AI implementation to ascertain the factors influencing employee trust.

It is likely that retraining will necessitate considerable effort, and older workers may be less inclined to participate. Consequently, it is vital to address any potential prejudice against AI among employees. From this standpoint, it is important to recognise that AI can offer opportunities and advantages. Nevertheless, any change requires guidance, and those occupying positions of technological leadership have a pivotal role to play in the AI revolution. Lessons learned from previous technology innovations (cloud, internet, etc.) are available, but are they enough? The accelerated pace of AI development presents a challenge in considering all relevant aspects (technical, ethical, legal, and reputational). A new risk landscape is emerging, training plans must be defined and implemented, technology functions and architectures need to be rethought and upgraded, human resource (HR) functions must be enhanced and renovated. As a result, the nature of decision-making is evolving, becoming increasingly structured and impactful. The HR function deals with a variety of demands and plays a crucial role. It is therefore a narrow perspective to view the role of HR as solely redeploying existing talent and/or attracting the right talent (which can be challenging in an increasingly competitive market). The HR function will be involved in setting the vision for the entire organisation, building learning and development capabilities through training programmes, evaluating the effectiveness of retraining, and ensuring that AI does not threaten employee well-being.

In general, the level of complexity will increase, and organisations may find themselves at turning points depending on the circumstances (e.g. large investments, unbundling of corporate function, business process outsourcing, staff reductions, etc.). Clear leadership and governance are essential, but do organisations already need new executives? A significant number of organisations are still considering where the responsibility for AI initiatives lies. This has resulted in the emergence of a multitude of roles and labels, which can be challenging for those seeking to gain an understanding of the optimal organisational structure. Examples of these include Chief Digital and AI Officer (CDAO), Chief AI Officer (CAIO), and Chief Information Security Officer (CISO), amongst others.

In the initial stages of AI implementation, a head of AI (not necessarily at the C-suite level) may be able to drive transformation, provided that he or she possesses both business and technical leadership skills and a comprehensive understanding of AI. This enables him or her to effectively assess the alternatives, strengths and weaknesses associated with AI. However, the number of tasks and responsibilities is increasing. Leaders must pursue continuous innovations and updates, evaluate investments, and maximise organisational benefits by reallocating resources, monitoring data and security breaches, assessing skill readiness, defining a talent strategy, optimising the use of AI, and so forth. Furthermore, they are involved in the establishment of rules,

policies, and processes to minimise the risk of harmful actions arising from the misuse of AI. This suggests that the role of leaders will become increasingly multifaceted.

However, is the personal and professional growth of leaders sufficient to address the evolving business landscape? It may not be sufficient to staff a new AI executive. In this new era, organisations need to evolve in order to fully leverage the potential of AI. To achieve this, it is recommended that a cohesive, organisation-wide approach be developed. In particular, organisations must commit to setting a vision, defining AI plans, and communicating the benefits and risks. By continuously monitoring the latter, they can be prepared and mitigate potential impacts before it is too late. As a result, organisations must build teams to manage complexity and increase collaboration and sharing in order to avoid making misguided decisions, be accountable, and provide an appropriate level of transparency to stakeholders.

References

Autor, D., Chin, C., Salomons, A. and Seegmiller, B. (2024) "New frontiers: The origins and content of new work, 1940–2018", *The Quarterly Journal of Economics*, Vol 139, No. 3, pp 1399-1465, https://doi.org/10.1093/qje/qjae008.

Bain & Company (2023) "Technology Report 2023", [online], https://www.bain.com/globalassets/noindex/2023/bain_report_technology_report_2023.pdf.

Daugherty, P., Burden, A. and Biltz, M. (2024) "Technology Vision 2024", [online], Accenture, https://www.accenture.com/content/dam/accenture/final/accenture-com/document-2/Accenture-Tech-Vision-2024.pdf#zoom=40.

Davis, F.D. (1989) "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS Quarterly*, Vol 13, No. 3, pp 319-340, https://doi.org/10.2307/249008.

Dutt, D., Ammanath, B., Perricos, C. and Sniderman, B. (2024) "Now decides next: Insights from the leading edge of generative AI adoption", [online] Deloitte, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-gen-ai-report.pdf.

European Commission (2018) "Artificial Intelligence for Europe", 25.4.2018 COM(2018) 237 final, [online], https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0237.

Goldman Sachs (2023a) "AI investment forecast to approach $200 billion globally by 2025", [online] https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html.

Goldman Sachs (2023b) "Navigating the AI Era", [online] https://www.goldmansachs.com/what-we-do/investment-banking/navigating-the-ai-era/multimedia/report.pdf.

Jyoti, R. and Schubmehl, D. (2023) "The Business Opportunity of AI", [online infographic], IDC Custom Solutions, https://news.microsoft.com/source/wp-content/uploads/2023/11/US51315823-IG-ADA.pdf.

Kelly, S., Kaye, S.A. and Oviedo-Trespalacios, O. (2023) "What factors contribute to the acceptance of artificial intelligence? A systematic review", *Telematics and Informatics*, Vol 77, 101925, https://doi.org/10.1016/j.tele.2022.101925.

McKinsey (2023) "The state of AI in 2023: Generative AI's breakout year", [online], https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year.

McKinsey (2024) "Beyond the hype: Capturing the potential of AI and gen AI in tech, media, and telecom", [online], https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/beyond-the-hype-capturing-the-potential-of-ai-and-gen-ai-in-tmt.

Mittal, N., Saif, I. and Ammanath, B. (2002) "Fueling the AI transformation: Four key actions powering widespread value from AI, right now", [online], Deloitte, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-state-of-ai-fifth-edition.pdf.

Shook, E. and Daugherty, P. (2024) "Work, workforce, workers: reinvented in the age of generative AI", [online], Accenture, https://www.accenture.com/content/dam/accenture/final/accenture-com/document-2/Accenture-Work-Can-Become-Era-Generative-AI.pdf#zoom=40.

Statista (2023) Artificial Intelligence: in-depth market analysis.

Wiles, J. (2023) "Take This View to Assess ROI for Generative AI", [online], Gartner, https://www.gartner.com/en/articles/take-this-view-to-assess-roi-for-generative-ai.

Generative AI Learning Environment for Non-Computer Science Engineering Students: Coding Versus Generative AI Prompting

Robert Alphinas and Torben Tambo
Department of Business Development and Technology, Denmark

roal@btech.au.dk
torbento@btech.au.dk

Abstract: This paper addresses the need for a positive and effective learning environment for engineering students in non-computer science fields to grasp Generative AI principles while navigating the intricate balance between its application and developmental insights. Drawing from pedagogical theories and cognitive science, especially Leinenbach and Corey (2004)'s Universal Design for Learning, this study proposes a framework tailored to the unique needs and backgrounds of engineering students. The framework emphasizes active learning strategies, collaborative problem-solving, and real-world applications to engage learners in meaningful experiences with Generative AI concepts. The central learning context is a M.Sc. program in management engineering with a course/training opportunity in Machine Learning Fundamentals using Python based on Google Collab. The introduction of Generative AI is based on selected Google libraries for Python. Furthermore, this paper explores various instructional approaches and tools to scaffold students' understanding of Generative AI, including hands-on projects, case studies, and interactive simulations. It also addresses ethical considerations and societal implications associated with Generative AI deployment, encouraging students to critically reflect on the broader impacts of their technical decisions. Through a synthesis of pedagogical best practices and AI development principles, this paper contributes to the ongoing discourse on effective AI education for non-computer science disciplines. By embracing a holistic approach that integrates theory with practical application, educators can empower engineering students to harness the transformative potential of Generative AI while navigating its complexities responsibly and ethically.

Keywords: Engineering education, Generative AI, Machine learning, Industrial engineering, Production engineering, Operations management

1. Introduction

The rise of artificial intelligence (AI) has fundamentally changed the game in various fields of engineering, going way beyond the usual boundaries of computer science (Brock & Von Wangenheim, 2019). As AI becomes more and more common, it's becoming increasingly important for engineering students in non-computer science disciplines to grasp AI concepts, especially Generative AI. These students need to navigate the complexities of AI development and application, which calls for a learning environment that supports both theoretical understanding, hands-on practice, and conversion to real-life settings (Yang, et al., 2020).

Generative AI, which is a branch of AI that focuses on creating new content like images, text, calculations, and music, has huge potential in a broad range of engineering fields (Radford, et al., 2019). But the ins and outs of developing and applying Generative AI models can be quite challenging, especially for students without a computer science background. That's why it's crucial to design educational frameworks that cater to their specific needs and make the most of their existing expertise (Wang & Shepherd, 2020).

This paper draws from pedagogical theories and cognitive science, particularly Leinenbach and Corey's (2004) Universal Design for Learning (UDL). UDL promotes flexible learning environments that accommodate the needs of diverse learners, engaging them through various ways of presenting information, taking action, expressing themselves, and getting involved (Meyer, et al., 2014). This approach is essential for creating a positive and effective learning environment for engineering students outside of computer science.

The main focus of this study is an MSc program in management engineering, which includes a course on Machine Learning Fundamentals using Python. To code and implement projects, students primarily use Google Colab, a user-friendly platform that allows them to experiment with Generative AI (Bisong, 2019). The introduction to Generative AI is facilitated through selected Google libraries for Python, enabling students to work with advanced AI tools and techniques without needing extensive prior knowledge of computer science (Chollet, 2021).

To help students grasp Generative AI concepts, this paper explores different teaching approaches and tools, including hands-on projects, case studies, and interactive simulations. These methods are designed to support learning step by step, giving students the opportunity to gradually build their knowledge and apply their skills in real-world situations (Chan & Lee, 2023). Additionally, the paper addresses the ethical considerations and

societal impacts associated with Generative AI, encouraging students to critically think about the broader effects of their technical decisions (Floridi, et al., 2021).

By combining best practices in teaching and principles of AI development, this paper aims to contribute to the discussion on effective AI education for non-computer science disciplines. By taking a holistic approach that integrates theory with practical application, educators can empower engineering students to harness the transformative power of Generative AI while responsibly and ethically navigating its complexities (Domingos, 2015).

2. Pedagogical Framework

The Universal Design for Learning (UDL) framework, developed by CAST (Center for Applied Special Technology), provides inclusive educational environments that work for all kinds of learners (Meyer, et al., 2014). It's based on three main principles (Figure 1) that address the different needs, preferences, and backgrounds of learners: Multiple Ways of Representing Information, Multiple Ways of Acting and Expressing Knowledge, and Multiple Ways of Engaging.

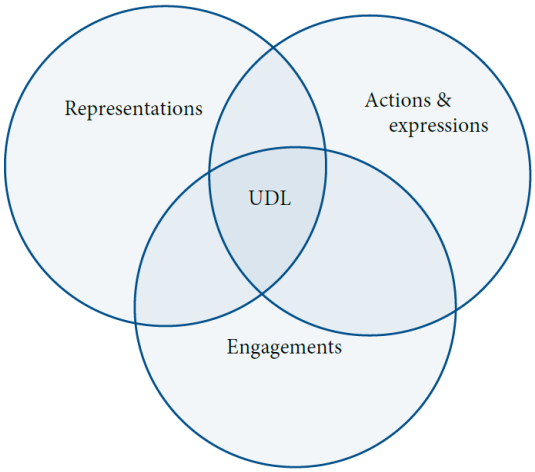


Figure 1: Pedagogical framework

Multiple Ways of Representing Information

This principle is all about giving learners different options for getting information and learning new things. Since people understand and process information in different ways, offering multiple ways of representing content makes sure that everyone can access it. Philippe et al (2020) found that using a variety of learning modes, like visual and auditory aids, can really improve understanding and retention, especially in STEM subjects (Philippe, et al., 2020).

Multiple Means of Action and Expression

This principle is all about giving learners different options to demonstrate their knowledge. We understand that everyone has their own way of expressing themselves, so it's important to offer a variety of ways to take action and show what they've learned. This includes (Novak, 2022): writing, speaking and hands-on activities such as projects, experiments, and activities that let learners apply their knowledge in real-life situations. A study by Tobin and Behling (2018) found that when students have different ways to take action and show what they know, their performance and engagement in higher education improve (Tobin & Behling, 2018). Brown and Knight (2022) also talked about how important it is to have diverse assessment strategies to encourage critical thinking and creativity in engineering students (Brown & Knight, 2022).

Multiple Means of Engagement

This principle aims to tap into learners' interests, challenge them appropriately, and motivate them to learn. Engagement can be achieved through (Meyer, et al., 2014) choice and autonomy, collaborative Learning and relevance and authenticity. Johansen et al. (2023) investigated the impact of choice and autonomy on student motivation and found significant improvements in engagement and learning outcomes (Johansen, et al., 2023). Additionally, Wiggins and McTighe (2019) highlighted the role of authentic learning experiences in enhancing student motivation and retention in STEM education (Wiggins & McTighe, 2019).

Application in Non-Computer Science Engineering Education

Multiple Means of Representation:

- To explain engineering concepts, use visualizations like neural network flow, 3D models, and simulations.
- When it comes to theoretical concepts, incorporate video lectures and podcasts.
- And for those who like to dig deeper, provide textual explanations and digital resources.

Multiple Means of Action and Expression:

- Let students show what they've learned through projects, lab work, and practical applications.
- Encourage them to present their ideas orally and have group discussions about engineering topics.
- Assess their understanding of engineering principles through written reports and design documentation.

Multiple Means of Engagement:

- Give students project choices that align with their interests and career goals.
- Foster collaboration on projects that simulate real-world engineering.
- Connect engineering problems to real-world applications, highlighting the importance and impact of the work.
- Implementing UDL in engineering education brings several benefits, including:
- Increased Accessibility: UDL ensures that all students, including those with disabilities, can access and engage with the material by providing multiple learning options.
- Enhanced Engagement: By using different teaching methods and learning activities, students stay motivated and interested.
- Improved Learning Outcomes: When education is tailored to meet the diverse needs of students, they develop a deeper understanding and retain knowledge better.

The UDL framework is essential for creating an inclusive and effective learning environment for engineering students. It's all about incorporating multiple means of representation, action and expression, and engagement. By doing so, educators can address the diverse needs of students and foster a more accessible, engaging, and successful learning experience.

3. Instrumental Approaches and Tools

Teaching Generative AI to non-software engineering students requires a multifaceted approach that includes:

- Hands-on projects
- Case studies
- Interactive simulations

Additionally, it is essential to incorporate ethical considerations and societal implications into the curriculum to prepare students for the broader impact of their technical decisions.

3.1 Hands-On Projects

Hands-on projects are an effective way to teach complex subjects like Generative AI. They allow students to apply theoretical knowledge to practical problems, enhancing their understanding and retention of the material (Kulesza, et al., 2015) , (Seo, et al., 2021).

- Step-by-Step Guidance: Providing detailed instructions for initial projects can help students build confidence. As their skills develop, they can take on more complex projects with less guidance (Chollet, 2021).
- Real-World Relevance: Projects should be designed to solve real-world problems relevant to the students' field of study. This relevance can increase motivation and demonstrate the practical utility of Generative AI (Wang & Shepherd, 2020).

Incremental Complexity: Starting with simple projects and gradually increasing the complexity can help students build their skills and avoid feeling overwhelmed (Prince, 2013).

Active Learning and Retention of Information

Studies have found that students who engage in active learning through hands-on projects retain information better and understand complex concepts more deeply compared to traditional lecture-based learning (Karpicke, 2011). Prince (2011) highlighted that active learning strategies significantly improve students' grasp of complex subjects. This includes hands-on projects. These methods enable students to apply concepts in real world scenarios.

Development of Essential Skills

Hands-on projects include problem-solving critical thinking and collaboration in the field of Generative AI (Kolodner, et al., 2011). Matar et al. (2024) discussed the integration of project-based learning in AI education. They achieve this while simultaneously building essential skills. Their findings suggest that students who engage in PBL are better prepared for the complexities of the AI field (Matar, et al., 2024). It demands critical evaluation of algorithms and their outputs. Collaborative efforts integrate AI solutions across various domains (Floridi, et al., 2021).

For instance, a project involving development of a Generative Adversarial Network (GAN) for creating synthetic images would require students to:

- -Solve technical problems related to model training. Optimize the model (problem-solving skills).
- -Critically evaluate the quality. Analyze ethical implications of generated images (critical thinking skills).
- Collaborate with peers. Integrate diverse expertise and perspectives (collaboration skills).

Students can work on projects that involve building generative models using frameworks like TensorFlow or PyTorch. For instance, creating Generative Adversarial Network (GAN) to generate images. Alternatively, they might build a variational autoencoder (VAE) for data compression. This presents practical challenges of working with large datasets (Chollet, 2021). Goodfellow et al. (2014) introduced GANs. Building models significantly enhances students' understanding of neural networks and deep learning principles (Chollet, 2021), (Goodfellow, et al., 2014).

3.2 Case Studies

The learning ladder of generative AI for management engineering students is proposed in Figure 2. To illustrate the learning ladder of generative AI for management engineering students, let's delve deeper into each stage of the process and this will help provide a comprehensive understanding of how each step builds on the previous one, forming a coherent path for integrating generative AI into business solutions.

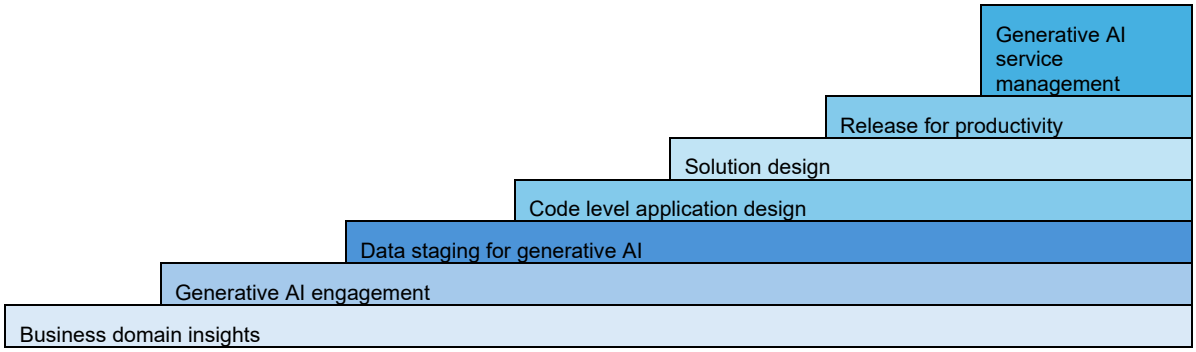


Figure 2: Learning ladder of Generative AI for management engineering students

Business Domain Insights: Understanding the business domain is crucial as it provides the context in which AI solutions will operate. This stage involves identifying business challenges and opportunities, setting objectives and understanding the nuances of industry (Sun, et al., 2018). Davenport and Ronanki (2018) explores how businesses are leveraging AI technologies to gain insights and improve decision-making processes.

Generative AI engagement: Engaging with generative AI involves selecting appropriate AI models, fine-tune them and understanding their capabilities. This stage sets the foundation for how AI can be utilized to address business needs (Brown, et al., 2020). Huang and Rust (2021) examines the impact of AI on service industries, with a focus on how AI technologies, including Generative AI, are used for market analysis and understanding customer needs (Huang & Rust, 2021).

Data staging for generative AI: Data staging involves preparing and preprocessing data to ensure it is suitable for AI model training and evaluation. It includes data cleaning, normalization and augmentation and splitting data into training and test sets.

Code level application design: This phase involves designing the application codebase including the architecture, APIs and integration with generative AI models. The focus is on modularity, scalability and maintainability.

Solution design: Solution design involves integrating AI models into full-fledged solution that meets business requirements. This includes designing the user interface, backend services, data pipelines, and deployment strategy.

Release for productivity: Releasing the solution involves deploying the application into production environment. This ensures it is reliable and scalable. It also must be maintainable. Continuous integration and continuous deployment (CI/CD) practices are critical here.

Generative AI service management: Service management is the process of monitoring and maintaining the deployed AI solution to ensure it still meets business needs. It involves: Performance monitoring, updating models, and handling user feedback.

Libraries

Google Colab is a cloud-based Jypyter notework environment and it supports Python and various popular libraries. It provides free access to GPUs and Tensor Processing Units (TPUs). TPUs are application-specific integrated circuits developed by Google specifically for accelerating machine learning workloads. They are designed to handle tensor operations. These operations are core part of the computations in machine learning models. This is particularly true for those involving deep learning. See Table 1.

Table 1: Key (generative) libraries in Google Colab

Libraries	Installation	Usage
TensorFlow and Keras	'!pip install tensorflow'	TensorFlow and Keras provide a robust framework for building and training neural networks, including generative models.
PyTorch	'!pip install torch torchvision'	PyTorch is another powerful library for deep learning, favored for its dynamic computation graph.
Hugging Face Transformers	'!pip install transformers'	Ideal for natural language processing tasks with pre-trained models like GPT-4.
Open AI API	'!pip install openai'	Access OpenAI's powerful models such as GPT-4 for text generation.
DALL-E-Mini	'!pip install dalle-mini'	Generate images from textual descriptions using DALL-E.
RunwayML	Varies based on the model; can be accessed via API or web interface.	Provides various generative models with a user-friendly interface

There are also other platforms which can provide the necessary computational resources, pre-built environments and integration with AI libraries to development and deployment of Generative AI models: Google Cloud AI Platform (Rajpurkar, et al., 2020), IBM Watson Studio, Paperspace Gradient, Run, Colab Pro+ etc.

3.3 Interactive Simulations

Students use insight in the industrial engineering field to capture the business problem. This gives the students ability to understand critical data, collect this with the appropriate context, and do the staging of data.

The generativity ensures that data a properly parsed. Algorithmically students can get insights from a statistical perspective with a minimal effort. This process can be elevated into an operating procedure in the production to secure precision, quality, reduce variability, and validate machining parameters. The students embeddedness in the production organization makes it accomplishable to implement the procedure for statistical process control and tooling. Here is an example of production line for wood furniture components with a simplified production line in Figure 3 and the data collection approach illustrated in Figure 4.



Figure 3:Simplified production line for wood analysis and cutting

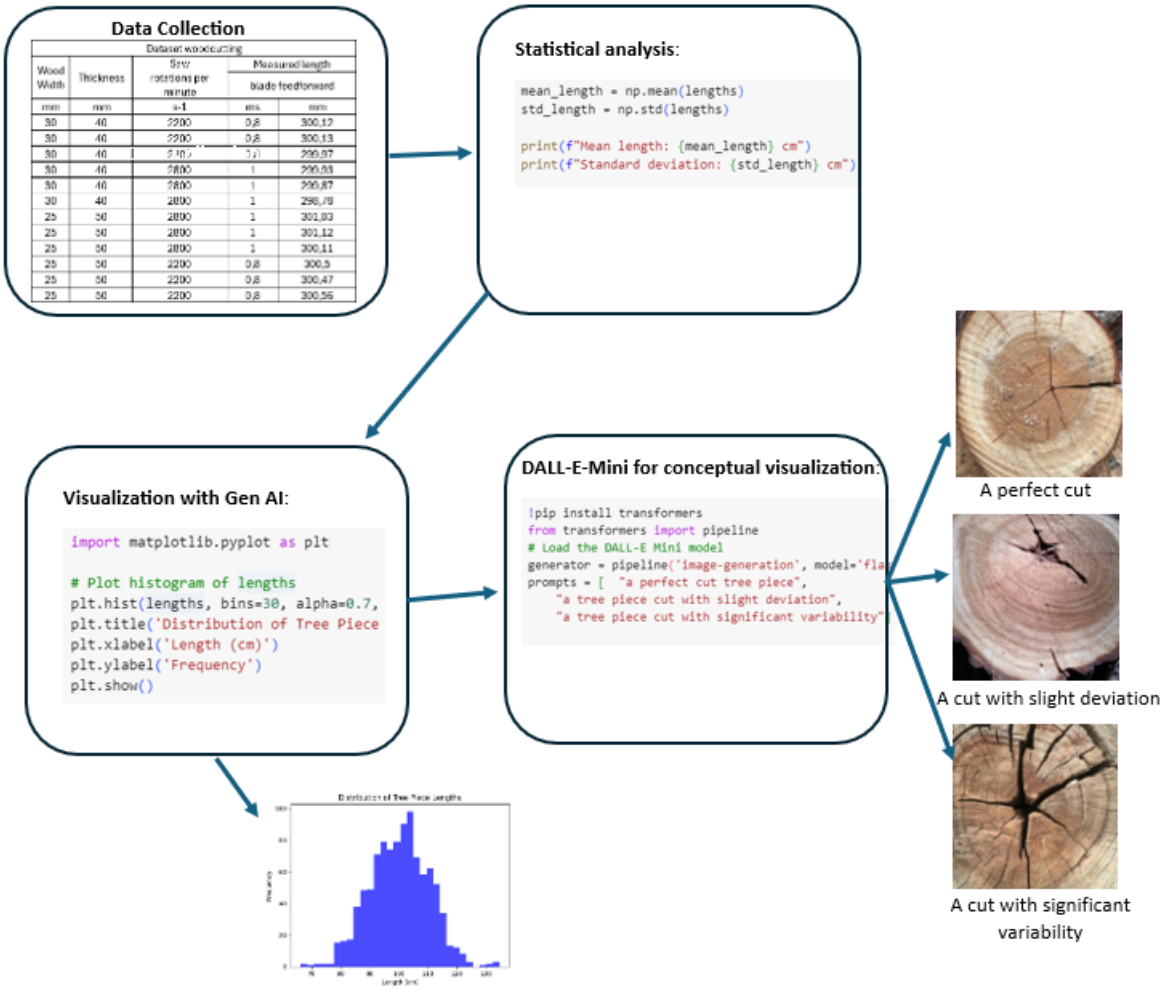


Figure 4: Generative thinking for digitalisation of wood end quality analysis

Analyzing the variability of cutting tree pieces with an automatic saw using generative AI thinking with a model like DALL-E-Mini involves a few conceptual steps. While DALL-E-Mini is designed for generating images from text prompts, the principles behind generative AI can help us think about how to analyze and visualize variability in a manufacturing process. Here's a conceptual approach.

- **Data Collection**: Gather data on the dimensions of the tree pieces cut by the automatic saw. This data should include measurements like length, width, thickness, and any other relevant parameters.

- **Statistical Analysis**: Use statistical methods to analyze the variability in the data. This could involve calculating the mean, standard deviation, and other measures of spread for each parameter.
- **Visualization**: Generate visual representations of the data to better understand variability. This is where generative AI can provide interesting insights by creating visual simulations or representations of the data distribution.

Implementation

- **Data Collection**: We generate a synthetic dataset of tree piece lengths using a normal distribution to simulate real-world variability.
- **Statistical Analysis**: Calculate mean and standard deviation to understand the central tendency and spread of the data.
- **Visualization with Generative AI Principles**: Use histograms to visualize the distribution of lengths. In a generative AI context, this step could involve creating synthetic images that represent the statistical properties of the dataset.
- **Using DALL-E-Mini**: Generate conceptual images based on text prompts that describe different cutting scenarios. These images can help visualize the variability qualitatively.

This approach combines traditional statistical analysis with the creative capabilities of generative AI to provide both quantitative and qualitative insights into the variability of tree piece cutting.

4. Framework

In management engineering, understanding the basics of machine learning (ML) is essential for applying generative AI in industrial contexts. Students without a software engineering background need to grasp core ML concepts, generative AI techniques, data preparation, model training, and evaluation. These foundational skills are vital for addressing industry-specific challenges and opportunities. To provide more specific context, specific generative AI methods, such as Generative Adversarial Networks **(**GANs**)** for creating realistic images and Variational Autoencoders **(**VAEs**)** for data compression and reconstruction, could be briefly mentioned. These methods are widely used in applications like synthetic data generation for testing business models or design optimization in manufacturing. Moreover, students will learn to manage generative AI services, focusing on ethical considerations and stakeholder engagement.

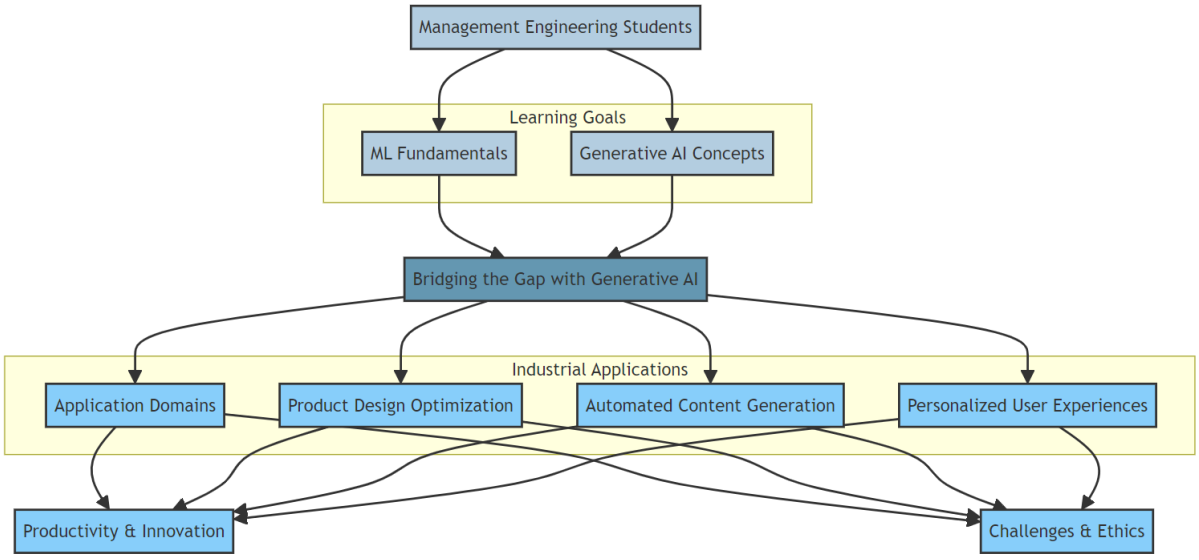


Figure 5: Framework of teaching generative AI for production engineers

Bridging the Gap: The connections between the education's learning goals and the industrial application steps using generative AI are shown here. For example, mathematical foundations lead to understanding business domain insights, algorithm understanding leads to generative AI engagement, and so on. Generative AI plays a pivotal role in bridging the gap between machine learning education and industrial applications. By integrating generative AI concepts and technologies into the curriculum, students gain practical experience that prepares them for real-world challenges. The scientific references provided illustrate how generative AI is being utilized

across various domains to drive innovation, enhance productivity, and manage AI services effectively. This holistic approach ensures that the theoretical knowledge gained in the classroom is directly applicable to solving practical business problems, thus closing the gap between academia and industry.

Generative AI techniques have revolutionized various industries by tackling complex problems such as automated content generation, product design optimization, and personalized user experiences. Generative AI optimizes product design processes by generating and evaluating design alternatives based on specified constraints and objectives. Bendoly et al. (2023) explores the integration of generative design techniques with additive manufacturing processes. It showcases how AI-driven generative design tools enable iterative optimization of product designs, leading to enhanced performance, reduced material waste, and accelerated innovation cycles (Bendoly, et al., 2023). Furthermore, Generative AI automates the creation of textual, visual, or multimedia content, enhancing efficiency and creativity in marketing, journalism, and entertainment. Benites et al. (2023) discusses advancements in automated text generation techniques using generative models like GPT (Generative Pre-trained Transformer) and their applications in content creation across various domains, highlighting improvements in natural language understanding and generation capabilities (Benites, et al., 2023). Jain and Singhal (2022) suggest surveys advanced recommendation systems powered by generative AI algorithms. They further discuss leveraging user data to generate personalized recommendations in e-commerce, streaming platforms, and digital marketing (Jain & Singhal, 2022).

5. Impact on Productivity and Innovation Vs Challenges and Ethical Considerations

Generative AI, especially in industrial contexts, shows immense promise in boosting productivity by automating intricate tasks and coming up with innovative solutions. For example, with Generative Adversarial Networks (GANs) used for synthetic sales data generation, businesses can now mimic various market conditions and customer behaviors without needing extensive real-world data collection. This means they can test business strategies and product offerings more quickly and iteratively, leading to a faster time-to-market and more informed decision-making (Karras, et al., 2019). For management engineering students, grasping and applying these AI-driven efficiencies can be transformative for business simulations, predictive analytics, and strategy development.

Generative AI is revolutionizing product design and development by enabling engineers and designers to explore a multitude of design possibilities beyond what humans could traditionally conceive. Tools like DALL-E and Stable Diffusion facilitate the creation of unique, high-quality visuals that can inspire new product concepts and marketing strategies (Ramesh, et al., 2021).

Generative AI excels at analyzing vast amounts of user data to create tailored content and recommendations, thereby enhancing personalized user experiences (Covington, et al., 2016). Moreover, AI-generated content, such as personalized emails, advertisements, and product suggestions, can significantly enhance customer engagement and drive sales.

Learning to apply AI in personalizing user experiences can help students develop customer-centric business models and strategies, ultimately improving business outcomes.

Challenges and Ethical Considerations

One of the foremost challenges in deploying generative AI is ensuring data privacy and security. These AI systems often require vast amounts of personal and sensitive information. Therefore, it's crucial to handle this data responsibly, anonymize it properly, and protect it from breaches. The General Data Protection Regulation (GDPR) and similar privacy laws impose strict requirements on data handling practices (Voigt & von dem Bussche, 2017). Management engineering students must be educated on the importance of data privacy and security.

Generative AI models can unintentionally perpetuate or amplify existing biases in their training data. For instance, if a GAN used in hiring processes is trained on biased data, it might produce biased hiring recommendations (Buolamwini & Gebru, 2018). Addressing these biases requires thorough auditing of training data and implementing fairness constraints in AI models. Students need to grasp the ethical implications of AI and learn techniques to identify and mitigate biases in AI systems, ensuring fairness and equity in AI-driven decisions. Another critical consideration is the ethical use of AI-generated content. AI can create highly realistic fake images, videos, and news, which can be misused to spread misinformation or manipulate public opinion (Chesney & Citron, 2019). Management engineering students should be trained to develop and adhere to ethical guidelines when using AI-generated content, promoting the responsible use of AI technologies. Ethical

challenges, such as algorithmic bias, data privacy concerns, and the broader societal impacts of AI, should also be emphasized as they are critical for stakeholder trust and engagement. For instance, bias in AI models can result in unfair treatment in areas like recruitment, while privacy issues can lead to misuse of sensitive data. Addressing these concerns requires interdisciplinary collaboration between engineers, ethicists, and policymakers to ensure that ethical challenges are tackled comprehensively.

The automation of tasks through generative AI can lead to job displacement, especially in roles involving routine and repetitive tasks. While AI can create new job opportunities and enhance productivity, it also requires a workforce skilled in AI and data science (Brynjolfsson & McAfee, 2014). Investing in education and training programs to reskill workers and prepare them for the AI-driven job market is crucial to mitigating the negative impact on employment. For students, this highlights the importance of continuous learning and skill development in AI and related fields to remain relevant in the evolving job market.

6. Conclusion

Generative AI offers significant potential for boosting productivity and fostering innovation across various industries. However, it also presents substantial challenges and ethical considerations that must be addressed. Generative AI's deployment in industries like healthcare (e.g., AI-assisted drug discovery) and manufacturing (e.g., generative design for optimized production) offers concrete examples where both its benefits and ethical concerns are at play. These examples illustrate how AI can boost productivity while also raising complex ethical dilemmas. The students can begin addressing these challenges through interdisciplinary learning, collaborating with stakeholders from diverse fields to ensure they are well-equipped to handle the ethical, technical, and societal implications of AI in their future careers. For management engineering students, learning to balance the benefits of AI with responsible and ethical practices is essential to ensure that AI deployment positively contributes to society. Future research and development should focus on enhancing the fairness, transparency, and accountability of AI systems while promoting inclusive and sustainable growth.

References

Bendoly, E. et al., 2023. The role of generative design and additive manufacturing capabilities in developing human–AI symbiosis: evidence from multiple case studies. *Decision Sciences,* Volume 00, pp. 1-21.

Benites, F., Benites, A. D. & Anson, C. M., 2023. Automated Text Generation and Summarization for Academic Writing. *Digital Writing Technologies in Higher Education,* pp. 279-301.

Bisong, E., 2019. *Building Machine learning and Deep learning models on Google cloud platform: A comprehensive guide for beginners.* Ottawa, Canada: Apress.

Brock, J. K. & Von Wangenheim, C. G., 2019. Demystifying AI for Managemant: An Exploratory Study. *Journal for Business Research,* Volume 102, pp. 75 - 83.

Brown, S. & Knight, P., 2022. Assessing learners in Higher Educationm. *Journal of Higher Education.*

Brown, T. B. et al., 2020. Language models are few-short learners. *Advances in neural information processing systems,* Volume 33, pp. 1877-1901.

Brynjolfsson, E. & McAfee, A., 2014. *The second machine age : work, progress, and prosperity in a time of brilliant technologies.* 1st ed. New York, N.Y.: W.W.: Norton & Company.

Buolamwini, J. & Gebru, T., 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.* s.l., FAT.

Chan, C. K. Y. & Lee, K. K. W., 2023. The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learningthan their Gen X and millennial generation teachers?. *Smart learning environments,* 10(1), pp. 60-83.

Chesney, R. & Citron, D., 2019. Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics. *Foreign Affairs (New York, N.Y.),* 98(1), pp. 147-155.

Chollet, F., 2021. *Deep learning with Python.* 2nd ed. USA: Manning publications Co..

Covington, P., Adams, J. & Sargin, E., 2016. *Deep Neural Networks for YouTube Recommendations.* Boston, Tenth ACM Conference on Recommender Systems.

Davenport, T. H. & Ronanki, R., 2018. Artificial intelligence for the real world. *harvard Business Review,* 96(1), pp. 108-116.

Domingos, P., 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world.* 1st ed. New York: Basic Books.

Floridi, L., Cowls, J., King, T. C. & Taddeo, M., 2021. How to design AI for social good: Seven essential factors. *Science and Engineering Ethics,* 24(6), pp. 1737-1762.

Goodfellow, I. J. et al., 2014. Generative adversarial nets. *International conference of neural information processing systems,* Volume 2, pp. 2672-2680.

Huang, M.-H. & Rust, R. T., 2021. Artificial intelligence in service. *Journal of service research,* 21(2), pp. 155-172.

Jain, A. & Singhal, A., 2022. *Personalized Food Recommendation—State of Art and Review.* Singapore, Springer.

Johansen, M. O., Eliassen, S. & Jeno, L. M., 2023. The bright and dark side of autonomy: How autonomy support and thwarting relate to student motivation and academic functioning. *The psychology of STEM Education, Frontiers in Education,* Volume 8.

Karpicke, J. D. J., 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science,* 331(6018), pp. 772-775.

Karras, T., Laine, S. & Aila, T., 2019. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Kolodner, J. L., Edelson, D. C. & Reiser, B. J., 2011. Project based inquiry science. *Science education,* 88(3), pp. 315-336.

Kulesza, T., Burnett, M., Wong, W. & Stumf, S., 2015. *Principles of explanatory debugging to personalize interactive machine learning.* New York, Association for computing machinery.

Matar, S. et al., 2024. *Cascade mentoring experience to engage high school learners in AI and Robotics through project-based learning.* Switzerland, Springer.

Meyer, A., Rose, D. H. & Gordon, D., 2014. *Universal design for learning: Theory and Practice.* Wakefield,Ma, USA: Cast professional Pub.

Novak, K., 2022. *UDL NOW! A teacher's guide to applying Universal Design for Learning.* 3rd ed. s.l.:CAST.

Philippe, S. et al., 2020. Multimodel teaching, learning and training in virtual reality: a review and case study. *Virtual Reality & Intelligent Hardware,* 2(5), pp. 421-442.

Prince, M., 2013. Does active learning work? A review of the research. *The research journal for engineering education,* 93(3), pp. 223-231.

Radford, A. et al., 2019. *Language Models are Unsupervised Multitask Learners.* s.l., Semantic scholar.

Rajpurkar, P. et al., 2020. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using Video pretraining. *Sceintific repots,* 10(1), p. 3958.

Ramesh, A. et al., 2021. *Zero-Shot text to image generation.* California, ArXiv.org.

Seo, K. et al., 2021. The impact of artificial intelligence on learner-instructor interaction in online learning. *International jounal of educational technology in higher education,* 54(18).

Sun, Z., Sun, L. & Stang, K. D., 2018. BigData analytics services for enhancing business intelligence. *Computational Information Systems,* Volume 58, pp. 162-169.

Tobin, T. J. & Behling, K. T., 2018. *Reach everyone, Teach everyone: Universal Design for Learning in Higher Education.* Morgantown WV: West Virginia University Press.

Voigt, P. & von dem Bussche, A., 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide.* 1st ed. Online: Springer International Publishing AG.

Wang, Y. & Shepherd, D., 2020. Exploring the Value of Generative Design in Engineering Education. *Journal of Engineering Education,* 109(4), pp. 614-639.

Wiggins, G. P. & McTighe, J., 2019. *Understanding by Design.* s.l.:ASCD.

Yang, Q., Steinfeld, A., Rosé, C. & Zimmerman, J., 2020. *Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design.* Honolulu, HI, USA, CHI Conference on Human Factors in Computing Systems, pp. 1-13.

An Exploratory Study on the Role of Generative AI in Designing Educational Board Games

Fabrizio Amarilli[1] and Oxana Timakova[2]
[1]Dublin City University, Ireland
[2]Civico Liceo Linguistico A. Manzoni, Milano, Italy

fabrizio.amarilli@dcu.ie
timakova.oksana@gmail.com

Abstract: The educational landscape is transforming due to new skill demands, fragmented attention spans, evolving communication patterns, and increased information access. Game-based learning (GBL) has emerged as a potential solution, fostering subject competency and broader skills. However, integrating games into education faces challenges for educators and schools. This study examines Generative AI's (GenAI) role in facilitating board game use in education, based on a case study in an Italian high school implementing a GenAI-enhanced board game for foreign language teaching. Two state-of-the-art GenAI technologies assisted in the whole GBL design process. Findings indicate GenAI can streamline board game integration through flexible rule adaptation, customized artifact production, and rapid content creation, reducing educator workload, as well as extend their capabilities with novel, creative solutions. Benefits are more pronounced when GenAI use is based on an overarching educational strategy. Study's findings have implications for educators and researchers seeking to effectively integrate GBL and GenAI in educational settings.

Keywords: Generative AI in education, Board game design, GBL, Language learning, AI-assisted pedagogy

1. Introduction

The 21st century has ushered in digitally-driven transformations that are altering human communication, cognition, and learning, particularly among young adults accustomed to instant messaging, fragmented attention spans, and ubiquitous access to knowledge (Prensky, 2012). This evolving context necessitates new skills emphasizing adaptability, rapid learning, and multitasking alongside deep subject-specific knowledge.

Game-based learning (GBL) has emerged as a promising approach to address these educational needs (Chou et al, 2023; Gao, 2023; Hanghøj, 2013; Prensky, 2012). Board games, in particular, offer advantages in accessibility and face-to-face interaction compared to digital alternatives (Bayeck, 2020; Gatti Junior et al, 2023; Luckner et al, 2024). However, integrating board games into education faces challenges, including inadequate training, fears of increased workload, rigid curricular requirements, and limited resources for customization (Cantoia et al, 2023; Hanghøj, 2013; York et al, 2022). Furthermore, the physical nature of board games can limit their adaptability compared to digital games. For instance, while a digital game might allow educators to easily switch between different maps or scenarios, a board game with a printed map or fixed components is inherently less flexible, potentially hindering alignment with specific learning objectives or adaptation to diverse educational contexts.

Generative Artificial Intelligence (GenAI) represents a potentially transformative solution, capable of streamlining and enhancing the entire GBL process. However, current literature addresses GenAI and GBL integration fragmentarily, focusing on specific design phases or privileging digital games (Fotaris et al, 2023; Gatti Junior et al, 2023; Tinterri et al, 2024). Research on GenAI applied to GBL, particularly board games, remains scant.

This study investigates how GenAI can be exploited in designing board games to foster GBL, analyzing a single case study of an Italian high school implementing a GenAI-enhanced board game for foreign language teaching. Through Participatory Action Research (PAR), we explore how GenAI can address challenges in integrating GBL and board games into educational practices.

Results indicate that GenAI can automate well-defined game design phases and extend educators' capabilities by suggesting novel ideas and creative solutions, with more evident impact when assisting throughout the entire design process. This study contributes to the dialogue on education's future, offering insights for leveraging GenAI for more effective, engaging, and adaptable learning experiences.

2. Theoretical Background

Game-based learning leverages games to facilitate learning, combining engaging gameplay with educational content (Perrotta et al, 2013; Plass et al, 2015; York et al, 2022). It enhances student engagement, promotes

active learning, and develops skills such as problem-solving and critical thinking. The generative potential of GenAI makes GBL a promising application domain (Gatti Junior et al, 2023; Mollick and Euchner, 2023; Tinterri et al, 2024).

GenAI and Large Language Models (LLMs) represent a transformative shift in AI capabilities, producing human-like outputs across various domains. Their impact extends to software development, content creation, data analysis, and process optimization, poised to transform task approaches and problem-solving across industries.

Emerging literature categorizes GenAI benefits into automation and augmentation (Mollick and Euchner, 2023). Automation optimizes well-defined tasks, improving efficiency, while augmentation expands workers' capabilities, enhancing output quality and creativity. Combined, these dimensions improve operational efficiency and empower workers to achieve higher performance levels.

Research on GenAI and GBL integration has explored various aspects, including assisting educators in game selection and customization (Tinterri et al, 2024), supporting instructional tasks in digital GBL (Yue and Guo, 2023), and theoretically proposing GenAI's potential for augmenting human capacities (Mollick and Euchner, 2023). However, much of the existing literature primarily emphasizes the automation capabilities of AI in educational contexts. This focus may overlook the broader potential of GenAI to augment pedagogical creativity and innovation.

Additionally, while digital games have been a primary focus of GBL research, there remains a significant gap regarding the application of GenAI to board games in educational contexts. Most studies concentrate on online or digital games, leveraging AI potential in code generation (Gozalo-Brizuela and Garrido-Mercha, 2023) and neglecting the unique advantages that board games offer in overcoming infrastructure limitations and facilitating face-to-face interactions (Luckner et al, 2024). This oversight limits our understanding of how board games can be effectively integrated into GBL frameworks.

Furthermore, literature on AI and GBL often addresses the use of AI in education by focusing on specific phases of the game design process. For instance, Hainey et al (2016) ’s systematic literature review highlights various isolated aspects of GBL implementation, while Gozalo-Brizuela and Garrido-Mercha (2023) explored GenAI's application in digital games, primarily analyzing specific design elements. Such an approach may restrict insights into how AI can enhance coherence throughout the entire learning experience. A more integrated perspective is necessary to fully understand the synergies between different design phases and their impact on GBL effectiveness.

3. Methodology

This study employs Participatory Action Research (PAR) to investigate GenAI's role in GBL adoption in education with a focus on board games. PAR enables collaborative knowledge creation between researchers and practitioners, fostering practical solutions to real-world educational challenges (Cornish et al, 2023). This methodology aligns with the study's aim to enhance educational practices through innovative technologies and pedagogical approaches.

The empirical knowledge base derives from a case study at an Italian high school between February and May 2024, focusing on 11 students, aged 18, learning Russian. The GBL endeavor aimed to improve language proficiency and students' ability to integrate knowledge from Russian literature and history courses.

The research utilized two prominent LLMs available at the time of the experimentation: OpenAI's ChatGPT (GPT-4) and Anthropic's Claude (3 Sonnet). The GBL design process was structured according to an adapted version from Nadwa Zulkifli et al (2023), comprising seven phases: Analysis of needs, Learning outcomes definition, Acceptance evidence, Game choice, Game personalization, Impact evaluation, and Instructional strategy.

For each phase, prompts were developed and submitted to both LLMs. Responses were cross-analyzed to identify contributions to the board game design process and highlight limitations. The assessment framework categorized contributions as either automation or augmentation (Mollick and Euchner, 2023). Automation focused on optimizing well-defined tasks, improving efficiency and resource utilization. Augmentation expanded educators' capabilities by providing additional knowledge, perspectives, and creative inputs. The evaluation of GenAI's contribution to board game design was conducted by an experienced educator who had previously implemented GBL in their classes. The educator's expertise allowed for a comparative assessment of GenAI’s impact against traditional GBL implementation methods used in previous years.

The prompt formulation followed an iterative approach, with initial submissions followed by analysis, revision, and resubmission as needed. Different strategies were tested, varying prompt formulation (open versus specific questions) and level of detail provided about students and educator choices throughout the GBL design process. The most effective approach emerged as a process where initial prompts were more general and open-ended, gradually becoming more detailed and analytical as the interaction with the LLMs progressed.

This methodology allowed for a comprehensive exploration of how GenAI can support educators in designing and implementing effective GBL experiences. The PAR approach ensured that the research remained grounded in practical educational needs while leveraging cutting-edge AI technologies to enhance pedagogical practices.

4. Findings

This study examines GenAI's role in designing the GBL process, focusing on teaching Russian in high school. Prompts for each GBL design phase were submitted to two LLMs, with responses analyzed by an experienced educator. The following sections present these prompts and analyze LLMs' outputs, offering insights into GenAI's potential in GBL, particularly for language education. The overall AI-based GBL process is presented in Figure 1. The figure highlights the specific contributions of GenAI and of the educator.

Figure 1: Phases and contributions in board game design

4.1 Design Phases and Prompts

The GBL design process was divided into seven phases, each supported by specific prompts directed at the two LLMs. An introductory prompt (P1) was prepared to provide to the LLMs the setting and the context of the design activity. The prompts of the different phases are highlighted in Table 1, while, in the following paragraphs we describe the rationale for the preparation of the prompts in each phase.

Table 1: Phases and prompts in board game design

P1	You are an experienced language teacher who is working with game-based learning (GBL). Your role is to design GBL lessons following a sequential model composed of the following phases: Analysis of needs (to understand students' characteristics that are relevant in GBL); Learning outcomes (LO) definition (to provide expected Los such as knowledge, skills, and attitudes); Assessment strategy (to identify quantitative and qualitative metrics to assess LO achievement); Game choice (to identify board games matching Los to be utilized); Game personalization (to turn Los into operation in a board game); Impact evaluation (to assess the impact of the activity for the customized game); Instructional strategy (to define the lesson plan for the GBL).
	Phase 1: Analysis of needs
P2	Let's start with the first step, Analysis of needs: Determine students' characteristics and needs. Create 10 questions to collect data on Interest, preference, player type of the students.
P3	Based on your suggestions I have revised the list of questions to collect the data from the students. I delivered the survey and collected responses from students belonging to one class. In the attached file you can find the question in the columns and the answers of the students in the rows. Can you analyze the answers to give suggestions on the creation of the board game for these students?
	Phase 2: Learning outcomes definition
P4	You are a teacher of Russian as a foreign language to this class of students. Students are studying the language, literature and history. However, students see these domains as independent. The overall objective is to help the students integrate the domains of language, literature and history. Students should be capable of positioning the literature authors into the correct historical context, so that the students can better understand and appreciate the authors' work. Outline 10 clear, measurable, and specific Los, in terms of knowledge, skills, and attitudes, that align with the overall objective.
	Phase 3: Assessment strategy
P5	Your aim is to use a board game to pursue the realization of the learning outcomes that you suggested. Now, the first step is to suggest quantitative and qualitative data that can be obtained to assess the Los before, during and after the gameplay.
	Phase 4: Game choice
P6	Suggest 5 off-the-shelf board games from which we can choose the most appropriate ones to pursue the achievements of the Los and the collection of data to assess the achievement of the LOs.
P7	To choose the game for the students, for each of the games that you suggested you need to prepare the following details: game name; game goal; rules; number of players; age rating; duration; game win rules; tools and artifacts; cooperation degree (individual, cooperative, collaborative, cooperative and collaborative); required language level (expected language proficiency of the players); Los (which Los you already suggested are expected to be pursued).
	Phase 5: Game personalization
P8	Based on games characteristics, I choose to play Timeline with the class. You need to change the rules of the game to make it collaborative rather than individual, as the students need to reinforce the collaboration transversal skill. Suggest 3 ways to revise rules to make Timeline a collaborative game.
P9	For Timeline with revised collaborative approach, prepare a description of game rules in Russian language B1 level (CERF classification of language knowledge).
P10	For Timeline with the revised collaborative approach, prepare a list of 20 historical events of Russian history. For each event provide title, date, and short description.
P11	For Timeline with the revised collaborative approach, prepare a list of 20 authors, writers and poets, in Russian literature.
P12*	For Timeline with the revised collaborative approach, and based on the first historical event, prepare one card that can be used in the game.
	Phase 6: Impact evaluation
P13	We will play Timeline with the first option of rules changes. For this, indicate the updated Los and qualitative and quantitative data that can be collected to assess the Los before, during, and after gameplay.
P14	Create a list of 15 questions, qualitative and quantitative, for pre-game phase to assess the following aspects. Quantitative: Pre-game quiz on Russian literary chronology and author-period associations; Self-assessment survey on students' perceived knowledge and collaboration skills. Qualitative: Short written responses on students' understanding of connections between Russian literature, language, and history.

P15	Create a list of 15 questions, qualitative and quantitative, for post-game phase to assess the following aspects. Quantitative: Post-game quiz to measure improvement in chronological understanding; Count of accurate connections made between authors, works, and historical events in a follow-up activity. Qualitative: Reflective essays on what students learned about Russian literary history and the benefits of collaborative learning.
P16	For post-game phase assessment, prepare a self-evaluation form, with 15 questions, which contains a survey on perceived improvement in knowledge and collaboration skills and a peer evaluation on team members' contributions and collaborative skills.
P17	For post-game phase assessment, create a list of 10 questions to be used in a focus group discussion on how the game changed students' perception of the interconnectedness of Russian literature, language, and history.
	Phase 7: Instructional strategy
P18	Based on the game selected, on the rules defined, on the assessment, pre, during, and post-game, create a lesson plan. Consider that one lesson is 50 minutes and the lesson plan can be implemented through more lessons.

* P12 was repeated for all suggested events.

Students' *Analysis of needs.* Two prompts guided data collection (P2) and analysis (P3) for student characterization. LLMs provided comprehensive questions covering game preferences, players' type, competitive spirit, and attitudes. The resulting survey yielded granular insights, informing game design aspects such as complexity and operational considerations. However, the LLMs also suggested collecting data on topics that students might be less equipped to self-evaluate accurately. For instance, ChatGPT proposed investigating students' perceptions of effective learning through games, while Claude suggested analyzing their comfort with making mistakes during the GBL process. While potentially valuable, these suggestions assume a level of metacognitive awareness and experience with GBL that students may not possess. Such assessments typically fall within the purview of experienced educators.

LLMs' suggestions were subsequently implemented in a survey administered to a class of 11 students. The resulting needs analysis was comprehensive and granular, revealing student preferences and informing various aspects of game design, including strategy complexity and operational considerations such as preferred game duration, frequency of play, and methods for rule clarification.

Two notable insights emerged from the LLMs' analysis. First, it was recommended to integrate both individual and collaborative elements into the game design. This approach aims to leverage students' existing preferences while simultaneously fostering the development of transversal competencies. Second, LLMs suggested implementing a progressive difficulty scale, which is a peculiarity of games, to enhance participant engagement.

Learning outcomes definition. A single prompt (P4) elicited suggestions for learning outcomes aligned with students' characteristics and educational objectives. The prompt was intentionally crafted to be both context-specific and open-ended. While the educator outlined the educational setting (a Russian as a foreign language class) and the multidisciplinary nature of the GBL (encompassing language, literature, and history), the request to LLMs was to identify an extended number of learning outcomes. This approach aimed to allow for a wide range of potential outcomes and to stimulate the LLMs to generate diverse and creative suggestions. The intentionally open-ended prompt resulted in a diverse set of outcomes spanning language proficiency, historical contextualization, critical thinking, and cultural awareness.

The LLMs demonstrated a nuanced understanding of the interdisciplinary nature of the task. They offered outcomes that connected the domains of language, literature, and history. With a diverse set of well-articulated outcomes, the LLMs enhanced the educator's ability to design a comprehensive and integrated learning experience. The range of suggestions provided flexibility in selecting the most suitable outcomes for the specific needs and abilities of the student group while aligning with the overall educational goals.

Assessment strategy. One prompt (P5) focused on identifying qualitative and quantitative data for outcome assessment. LLMs suggested methods such as pre- and post-game tests, observational assessments, and reflective journaling, along with operational guidance on data collection. The breadth of suggestions highlights the LLMs' potential to enhance educators' capabilities in designing evaluation strategies. However, the final selection and implementation depend on the educator's professional judgment to tailor methods to specific learning outcomes, student characteristics, and practical constraints.

Game choice. Two prompts (P6, P7) guided the selection of an appropriate off-the-shelf board game. LLMs provided game suggestions with detailed descriptions, including information about game mechanics and educational potential, as well as practical suggestions for implementation. Both responses were analytical overviews that closely matched the prompt requirements. The responses provided enough detail to help the educator make an informed decision. Importantly, some suggested games were unfamiliar to the educator, demonstrating that the LLMs could broaden their knowledge and expand their options.

Game personalization was implemented through five prompts (P8-P12), demonstrating a progressive refinement of the game design process with increasing educator involvement. P8 asked for the adaptation of the Timeline game to a collaborative format, reflecting the educator's unique knowledge of the class needs, particularly the importance of fostering collaborative skills—information not available to the LLMs. Both LLMs provided focused and creative responses, demonstrating their ability to adapt game mechanics to specific educational requirements. P9 requested game rules in Russian at a B1 language level, emphasizing the production of specific game artifacts. While both LLMs provided effective responses in Russian, analysis using Textometer (Dmitrieva, 2023; Laposhina et al, 2018) showed that ChatGPT's output was at a B2 level, indicating some inconsistency in adhering to the language level. P10 and P11 focused on generating additional game materials, specifically historical events and literary authors. Although the LLMs provided correct responses, this stage emphasized the crucial role of the educator in aligning content with the school curriculum—a context beyond the LLMs' knowledge. P12 asked for the production of card materials based on the educator's selections. This prompt revealed differences in LLM capabilities: ChatGPT generated images, while Claude provided extended descriptive text. For example, for the "Baptism of Russia" event, ChatGPT produced an effective image, while Claude offered a comprehensive historical and cultural context. This complementarity in LLM functionalities proved valuable for creating rich educational materials. However, the image generation also exposed potential biases in LLM training data, as seen in an anachronistic depiction of the 1957 Sputnik satellite. This phase highlighted the complementary strengths of different LLMs In game modification and material creation while also revealing challenges in language level consistency and potential biases in generated content.

Impact evaluation. Five prompts (P13-P17) addressed various aspects of evaluating the game-based training activity, focusing on the specific personalized game. LLMs identified updated learning outcomes specific to the adapted collaborative game and determined appropriate data collection methods (response to P13). They provided specific and operational contributions, detailing not just what to assess but also how to collect the data. Notably, Claude emphasized the interdisciplinary connections between language, literature, and history, aligning closely with the educator's original objectives. LLMs proposed precise, analytical, and operational questions for pre- and post-gameplay assessment of the students (response to P14 and P15). In suggesting self-evaluation questions (P16), LLMs also proposed cross-evaluation methods, such as having students assess their teammates' collaborative spirit and leadership qualities, adding a valuable peer assessment dimension to the evaluation process. In the identification of questions for a post-gameplay focus group (response to P17), LLM' suggestions extended beyond assessing knowledge gains, incorporating questions about how the game experience altered students' perspectives on the subject matter. This approach offered the educator an opportunity to explore deeper, more transformative learning outcomes. Throughout this phase, the LLMs demonstrated their capacity to augment the educator's evaluation strategies, providing a rich array of assessment tools and perspectives.

Instructional strategy. The final phase focused on the definition of a specific lesson plan (P18). This phase integrated the collaborative-oriented rules, specific learning outcomes, and assessment approaches developed in earlier stages. The responses provided by the LLMs demonstrated a high level of effectiveness in capturing and synthesizing the educator's requirements. Furthermore, the lesson plans incorporated key suggestions from earlier phases, such as providing the educator with elements on how to implement a progressive difficulty curve to enhance student engagement. This showcased LLMs' ability to maintain consistency and coherence throughout the design process.

LLMs demonstrated their potential to both optimize and extend the educators' capabilities, streamlining activities, offering diverse perspectives, creative solutions, and detailed operational guidance. However, the interplay with the LLMs unveiled the limitations of the technology and underscored the crucial role of educators in guiding the design process, ensuring curriculum alignment, and critically evaluating LLM outputs.

4.2 Analysis of the LLM contribution to GBL design

The integration of LLM in GBL design demonstrated significant impacts in terms of both automation and augmentation. Automation benefits primarily manifested as increased efficiency, reduced workload for educators, and streamlined processes. These included rapid generation of assessment tools, swift creation of lesson plans, and efficient production of game materials. Augmentation benefits expanded educators' capabilities by providing novel ideas, diverse perspectives, and creative solutions. This included suggesting innovative game adaptations, offering interdisciplinary learning outcomes, and introducing educators to new game options. The synergy between these two aspects of LLM contribution proved particularly valuable in enhancing the overall design process. Table 2 presents a detailed breakdown of the LLMs' contributions across the seven phases of the game design process and the role of the educator.

Table 2: Benefits and limitations of GenAI in the board game design process

Design phases	Automation benefits	Augmentation benefits	Educator's role
1. Analysis of Needs	Rapid generation of student assessment questions	Comprehensive perspective on student characteristics to consider	Selection and adaptation of questions based on specific class context
2. Learning outcomes definition	Quick generation of diverse learning objectives	Creative suggestions for interdisciplinary outcomes	Refinement and alignment with curriculum standards
3. Assessment strategy	Efficient creation of assessment metrics	Innovative ideas for qualitative and quantitative measures	Selection and customization of assessment methods
4. Game choice	Rapid suggestion of multiple game options	Introduction to new, potentially unknown games	Final selection based on class needs and resources
5. Game personalization	Swift generation of game adaptations and materials	Creative ideas for game modifications and new content	Guiding the personalization process with class-specific insights
6. Impact evaluation	Quick development of diverse evaluation tools	Suggestions for multi-faceted assessment approaches	Tailoring evaluation methods to class objectives and student needs
7. Instructional strategy	Rapid creation of lesson plan templates	Innovative ideas for integrating the game into broader curriculum	Adapting and finalizing the instructional approach

In the needs' analysis, LLMs demonstrated strong automation capabilities by quickly generating comprehensive sets of student assessment questions, significantly reducing the time educators needed to develop these tools. The LLMs provided a wide range of questions covering various aspects of student characteristics, including language proficiency, learning preferences, and prior knowledge of Russian literature and history. The augmentation aspect was also present, as LLMs offered diverse perspectives on student characteristics, potentially broadening the educator's view.

The Learning outcomes definition phase shifted towards augmentation, with LLMs providing creative and interdisciplinary learning objectives. This expanded the educator's perspective and facilitated a more holistic approach to language learning integrated with literature and history. The LLMs suggested outcomes that addressed not only language proficiency but also cultural understanding, critical thinking, and historical context analysis.

For the Assessment strategy, LLMs efficiently automated the creation of assessment metrics while also augmenting the process with innovative ideas for both qualitative and quantitative measures. LLMs proposed a mix of traditional language assessments, creative writing tasks, and collaborative projects, as well as peer evaluation methods and reflective journaling.

The Game choice phase strongly emphasized augmentation, as LLMs introduced educators to potentially unknown games and offered creative suggestions for game selection. This significantly expanded the educator's repertoire of educational game options. The LLMs provided detailed descriptions of various board games and offered insights into how each game could be modified to incorporate elements of Russian literature and history.

In Game personalization, LLMs exhibited strong augmentation capabilities by providing creative ideas for game modifications and generating new content. This phase also saw automation benefits in the swift production of adapted game materials. The LLMs suggested innovative ways to incorporate Russian literary quotes, historical events, and cultural references into the game mechanics and rapidly generated custom card content, game board designs, and rule modifications.

The Impact evaluation phase balanced automation and augmentation, with LLMs quickly developing diverse evaluation tools while also suggesting multi-faceted assessment approaches. The LLMs proposed a combination of pre- and post-game assessments, in-game performance metrics, and reflective exercises to comprehensively evaluate students' progress.

Finally, the Instructional strategy phase leaned more towards automation, with LLMs rapidly creating lesson plan templates, though still offering some augmentation through innovative ideas for curriculum integration. LLMs provided detailed lesson structures and suggested ways to connect GBL experience with regular classroom activities and homework. The analysis revealed a pattern where automation benefits were more pronounced in the initial and final phases of the design process, while augmentation dominated the middle phases, where creative input and novel ideas were most valuable.

The application of LLMs across the entire design process generated benefits that transcended the contributions to individual phases. This holistic approach created a synergy between different stages of the design, enhancing the overall coherence and effectiveness of the GBL experience. For instance, the evaluation strategies developed in the final phase were directly informed by and aligned with the acceptance evidence designed at the beginning of the process. Similarly, the lesson plan design incorporated suggestions for progressive difficulty that emerged during the initial needs analysis phase.

Throughout this process, the educator's role remained crucial. While LLMs provided valuable suggestions and efficient tools, the educator was responsible for critical decision-making, contextualizing LLM’s outputs to the specific classroom environment, and ensuring alignment with curriculum standards and pedagogical goals. The educator's expertise was particularly vital in selecting appropriate assessment questions, refining learning outcomes, choosing suitable games, guiding personalization, and adapting evaluation methods and instructional strategies to meet the students’ unique needs.

This synergy between LLM’s capabilities and educator expertise demonstrates the AI potential to enhance the GBL design process while underscoring the irreplaceable role of human judgment and pedagogical experience in effective educational design. The complementary nature of AI assistance and human expertise resulted in a richer, more innovative, and more tailored educational experience than either could have achieved alone.

5. Discussion and Conclusion

This study explored GenAI's role in GBL, focusing on integrating language, literature, and history in Russian language learning and demonstrated AI’s potential to automate and augment the design process. The findings contribute to existing literature on AI in education in several ways.

First, while much extant literature focuses on AI's automation capabilities (Tinterri et al, 2024), our research highlights the substantial augmentation potential of GenAI. This shift addresses a limitation in the literature that often overlooks how AI can enhance pedagogical creativity and innovation. Our results show that AI can not only streamline routine tasks but also extend educators' capabilities by providing novel ideas, diverse perspectives, and creative solutions. This augmentation was particularly evident in the middle phases of the game design process, expanding our understanding of AI's role in education beyond efficiency gains to a tool for promoting educational creativity and innovation.

Second, this study contributes by focusing on the application of AI in educational board game design, in contrast to existing literature primarily addressing online or digital games. This focus is particularly relevant given board games' accessibility in various educational settings, including those with limited technological infrastructure. By demonstrating how AI can support board game design and adaptation, this study opens new possibilities for implementing GBL without extensive digital resources.

Third, our research adopts a holistic approach to AI-assisted GBL, examining AI application across the entire GBL development spectrum. This integrated perspective responds to the fragmented approach observed in existing literature, which often analyzes AI role within isolated phases of the game design process. By revealing synergies between different phases, we provide a more comprehensive understanding of how AI can enhance

coherence and effectiveness throughout the entire GBL experience. Furthermore, the study unveils the complementarity and the added value created through the integration of different AI technologies.

The observed pattern of AI's contributions suggests a strategic approach to AI integration. Educators and instructional designers can leverage AI's efficiency-enhancing capabilities in the planning and implementation stages while embracing its creative potential during the core design phases.

The study underscores the critical role of human expertise in the AI-assisted design process. While AI provided valuable suggestions and efficient tools, the educator's judgment remained crucial in contextualizing AI outputs, ensuring curriculum alignment, and tailoring the game to specific student needs. This finding contributes to the ongoing discourse on human-AI collaboration in education, emphasizing the complementary nature of AI capabilities and human pedagogical expertise.

Despite these contributions, the study shows limitations that represent opportunities for future research. The focus on a single case study in foreign language teaching limits generalizability to other educational contexts. This localized setting may not reflect the diverse challenges and dynamics present in different regions or educational systems. The study concentrated on language teaching, specifically examining aspects such as communicative competence and capacity to link foreign language proficiency, literature, and history. These specificities may not translate directly to other subjects or educational levels. Future research should extend this approach to diverse educational contexts and larger sample sizes to better understand the broader applicability of GenAI in GBL design process.

Furthermore, the study investigated the role of AI in GBL within a limited timeframe. Future studies could adopt a longitudinal approach and explore long-term impacts of AI-designed games on student learning outcomes, motivation, and engagement. Additionally, the complementary strengths of various LLMs demonstrate potential synergies in combining multiple AI tools, opening an avenue for future research.

Finally, research into educators' professional development needs to effectively leverage AI in game design would be valuable. Investigating ethical implications of AI use in educational design, including data privacy, algorithmic bias, and the changing role of educators, represents another area for future inquiry.

References

Bayeck, R. Y. (2020) "Examining board gameplay and learning: A multidisciplinary review of recent research", *Simulation & Gaming*, Vol 51, No. 4, pp 411-431.

Cantoia M., Clegg A., Tinterri A. (2023) "Training teachers to design game-based learning activities: evidence from a pilot project", *Computers in the Schools*, pp 1-23.

Chou, Y. S., Hou, H. T., Chang, K. E., and Su, C. L. (2023) "Designing cognitive-based game mechanisms for mobile educational games to promote cognitive thinking: an analysis of flow state and game-based learning behavioral patterns", *Interactive Learning Environments*, Vol *31*, No. 5, pp 3285-3302.

Cornish, F., Breton, N., Moreno-Tabarez, U., Delgado, J., Rua, M., de-Graft Aikins, A. and Hodgetts, D., (2023) "Participatory action research", *Nature Reviews Methods Primers*, Vol *3,* No. 1, p.34.

Fotaris, P., Mastoras, T., and Lameras, P. (2023) "Designing educational escape rooms with generative AI: A framework and ChatGPT PROMPT engineering guide", *Proceedings of the European Conference on GBL*.

Gao, L. (2023) "Literature Review for the Design Methods of Serious Games", *Proceedings of the 17th European Conference on GBL: ECGBL 2023*. Academic Conferences and publishing limited.

Gozalo-Brizuela, R. and Garrido-Merchán, E.C. (2023) A survey of Generative AI Applications. *arXiv preprint arXiv:2306.02781*.

Gatti Junior, W.G., Marasco, E., Kim, B., Behjat, L. and Eggermont, M. (2023) "How ChatGPT can inspire and improve serious board game design", *International Journal of Serious Games*, Vol 10, No. 4, pp 33-54.

Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., and Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Computers & Education*, *102*, pp. 202-223.

Hanghøj T. (2013) "Game-based teaching: Practices, roles, and pedagogies", In De Freitas. S., Ott M., Popescu M.M. (Ed) *New Pedagogical Approaches in Game-enhanced Learning: Curriculum integration*, IGI Global

Luckner, N., Pollak, M., and Purgathofer, P. (2024, June). Exploring the Impact of Purposeful Board Games in Higher Education. In *International Conf. on Human-Computer Interaction*. Spr. Nature Switzerland, pp. 69-81.

Mollick, E. and Euchner, J. (2023), "The Transformative Potential of Generative AI: A Conversation with Ethan Mollick", *Research-Technology Management*, Vol 66, No. 4, pp 11-16.

Perrotta, C., Featherstone, G., Aston, H. and Houghton, E. (2013), *Game-based Learning: Latest Evidence and Future Directions* (NFER Research Programme: Innovation in Education). Slough: NFER

Plass, J. L., Homer, B. D., and Kinzer, C. K. (2015) Foundations of game-based learning. *Educational psychologist*, Vol 50*,* No. 4, pp 258-283.

Prensky, M. R. (2012) *From digital natives to digital wisdom: Hopeful essays for 21st century learning*. Corwin Press, California.

Tinterri, A., Di Padova, M., Palladino, F., Vignoli, G., and Dipace, A. (2024), "AI in board Game-Based Learning", *CEUR Workshop Proceedings*.

Zulkifli, N. N., Zhang, Y., Ayub, A. F. M., and Nawi, N. R. C. (2023) "From Analysis to Creation: Utilizing the ADDIE Model for Developing and Educational Game for Children", *Proceedings of the Asian Conference on Education 2023*.

York, J., deHaan, J., Childs, M., and Collins, M. (2022) "How is gamification like being trapped in the Matrix? And what is the "real-world" of game-based learning", *Digital Culture & Education*, Vol 14, No. 3, pp 35-54.

Yue L. and Guo L. (2023) "Combine DGBL With AI System: A Technical Guidance to Reduce Teacher's Burden in Digital Game-Based Learning", *In European Conference on GBL*. Academic Conf. Int. Limited, pp 826-XXV.

Artificial Intelligence Life Cycle: The Detection and Mitigation of Bias

Ashionye Aninze and Jagdev Bhogal
Birmingham City University, UK

ashionye.aninze@mail.bcu.ac.uk
jagdev.bhogal@bcu.ac.uk

Abstract: The rapid expansion of Artificial Intelligence(AI) has outpaced the development of ethical guidelines and regulations, raising concerns about the potential for bias in AI systems. These biases in AI can manifest in real-world applications leading to unfair or discriminatory outcomes in areas like job hiring, loan approvals or criminal justice predictions. For example, a biased AI model used for loan prediction may deny loans to qualified applicants based on demographic factors such as race or gender. This paper investigates the presence and mitigation of bias in Machine Learning(ML) models trained on the Adult Census Income dataset, known to have limitations in gender and race. Through comprehensive data analysis, focusing on sensitive attributes like gender, race and relationship status, this research sheds light on complex relationships between societal biases and algorithmic outcomes and how societal biases can be rooted and amplified by ML algorithms. Utilising fairness metrics like demographic parity(DP) and equalised odds(EO), this paper quantifies the impact of bias on model predictions. The results demonstrated that biased datasets often lead to biased models even after applying pre-processing techniques. The effectiveness of mitigation techniques such as reweighting(Exponential Gradient(EG)) to reduce disparities was examined, resulting in a measurable reduction in bias disparities. However, these improvements came with trade-offs in accuracy and sometimes in other fairness metrics, identified the complex nature of bias mitigation and the need for precise consideration of ethical implications. The findings of this research highlight the critical importance of addressing bias at all stages of the AI life cycle, from data collection to model deployment. The limitation of this research, especially the use of EG, demonstrates the need for further development of bias mitigation techniques that can address complex relationships while maintaining accuracy. This paper concludes with recommendations for best practices in Artificial Intelligence development, emphasising the need for ongoing research and collaboration to mitigate bias by prioritising ethical considerations, transparency, explainability, and accountability to ensure fairness in AI systems.

Keywords: Artificial intelligence, Machine learning, Ethics in AI, Data bias, Machine learning fairness, Bias in AI

1. Introduction

This paper aims to raise awareness of the increasing importance of mitigating biases within Machine Learning(ML) algorithms, detecting them, and effectively reducing them. We interact with Artificial intelligence(AI) systems without being aware. It is in our home devices, banking systems, search engines, and any system that requires intelligence. As AI is increasingly becoming a part of our daily lives, these systems must be fair, reliable, and trustworthy. The issue regarding the fast expansion of AI is that there are not many rules or regulations regarding ethics considered or developed. While investment schemes run by(GOV.UK, 2023), have been set up and AI bills are in the process of being made(Parliament.uk, 2023), those that do exist are in their early days. There is a gap regarding ethics in AI and without a deeper understanding of how our own biases can be transferred into the algorithms we create; biased AI systems will continue to be developed. This could have many repercussions socially and legally such as providing harmful information or reinforcing biased opinions. AI can have a negative influence and there should be more guidance for developers to effectively address bias.

1.1 Bias in AI

According to Google Cloud(2024), AI is a broad field referring to technologies that build machines and computers that can replicate cognitive functions associated with human intelligence. This includes seeing, understanding, and responding to written or spoken languages. AI can analyse data, and make predictions, often thought of as a system - AI is a set of technologies implemented in systems to learn and solve complex problems.

ML is a subset of AI that enables a machine or system to learn and improve from experience. ML algorithms improve their performance over time as they are trained on more data. ML is the application of AI that allows machines and systems to extract knowledge from the data and learn by itself. Further stated by Google Cloud(2024), AI Is the idea that a machine can mimic human intelligence, but ML does not. ML aims to teach machines how to perform a task based on patterns provided by data.

Bias in AI refers to the systematic and structured errors in AI systems that result in unfair outcomes and inequalities. As ML is becoming widely used, biases within these decision-making machines must not replicate the biases that exist within training data. Park and Hu(2023) state that biases can arise at various stages of AI

development, including data-generating processes and human decision-making. As ML algorithms learn from data if the data given is biased, the algorithm will learn from that and continue perpetuating the biases within that dataset.

1.2 Prevalence of Bias in AI and Potential Consequences

In recent years, the growing implications of AI have garnered significant attention within the scientific community and our society. As stated by Rishi Sunak during the AI safety summit in November 2023 the UK will be taking unique steps to establish the first AI safety Institute to 'Understand what each new model is capable of, exploring all the risk from social harms like bias and misinformation'(The Telegraph, 2023). Sunak's announcement highlights the growing prevalence of biases within AI systems, prompting nations to address these challenges and mitigate their impact collaboratively. This highlights that biased AI algorithms are becoming an international issue and therefore must be addressed. Bias is the act of supporting or opposing a particular thing or person unjustly because of personal opinions that influence judgment. In AI and ML, bias refers to the tendency of algorithms to produce results that favour certain outcomes or groups of people over others, this is due to the underlying human bias or training data(Chopra and Khurana, 2023). This can lead to unfair outcomes or inaccurate predictions, especially in sensitive industries like healthcare or justice. The outcomes often disadvantage specific groups of people, especially those from marginalised communities, reinforcing existing societal inequalities(Akselrod, 2021). There is ample evidence of discriminatory harm that AI tools can cause to already marginalised groups, Akselrod(2021) explains that AI is built by humans and deployed within systems that have been established with discrimination – from housing to the workplace. The bias preexisting in the data that are used to train the AI is often discriminatory or under-representative for people of colour, women, and any other marginalised group. Only those who understand can work to address the potential harm.

2. Literature Review

This literature review examines multiple topics and themes relevant to ethics within AI. In a comprehensive exploration of bias in AI algorithms, this report will address the ethical considerations within ML highlighting real-world consequences.

Bias mitigation between high-income and low-middle income is explored in a paper by Yang et al.(2024). Within their research, they revealed the preexisting biases occurring within their dataset that were further manifested with their Neural Network Model. Conclusions revealed OD disparities were reduced at the expense of model accuracy. In another paper by González-Sendino, Serrano and Bajo(2024), they explore several datasets. This includes the adult dataset. The paper goes beyond observing correlations and takes a causal model approach, which aims to represent a cause-and-effect relationship between the different variables.

2.1 Different Variations of Bias in AI

Within the context of AI, Ferrara(2023) states that bias can arise from various sources of the ML pipeline. ML models can learn and replicate patterns of bias present in the data used to train them, resulting in unfair or discriminatory outcomes. Algorithmic bias refers to the systematic and structured errors in an AI system that generates unfair results. It can emerge from intentional design choices or unexpected decisions associated with how data is gathered(Shin and Shin, 2023). Understanding and addressing algorithmic bias is fundamental for ensuring fairness and equality in decision-making processes, biased algorithms can lead to discriminatory outcomes, perpetuating existing inequalities. Data bias occurs when data used to train or test a model is not representative of the real-world population, or when certain groups are over or under-represented(Wamburu et al., 2022). This can happen when the data is collected from biased sources. Social bias manifests when individuals or groups tend to favour or discriminate against social characteristics, such as race, gender, or socioeconomic status. It involves the formation of stereotypes, prejudices, and discriminatory behaviours(Masha Fedzechkina, Lucy Hall Hartley, and Roberts, 2022.

The development of fairer and more ethical AI systems can be developed by addressing diverse types of biases within the AI lifecycle.

2.2 Examples of Bias in AI

The problem of bias is shown in a paper by Raji and Buolamwini(2019) which found statistics that facial recognition algorithms used by law enforcement are up to 100 times more likely to misidentify Black or Asian faces compared to White faces. Algorithms like this could lead to the misidentification of Black or Asian individuals, resulting in wrong arrests and allowing actual perpetrators to avoid justice.

Nazer et al.(2023) states, 'Race, sex and disability status are inequities that may determine which health-related problems are prioritised and funded.' This decides what research, including those related to AI, is produced. Biases would result in research questions and problems favouring a segment of the population regardless of the burden of the disease. This can lead to unbalanced research in certain demographics leading to skewed data which produces skewed algorithms. This draws attention to the importance of fair and balanced research questions, as the failure to address such biases could perpetuate health disparities as the focus may be disproportionally on diseases affecting certain ethnic groups.

3. Methodology

This section provides an overview of the steps to detect and mitigate biases within the ML model and dataset.

The data mining methodology follows the KKD process referred to in a paper by Zavaleta-Sánchez et al.(2024) excluding the closing section of deployment. The Adult dataset Becker & Kohavi(1996) was selected for its popularity in fairness and bias research due to its demographic attributes and imbalanced income target(reflecting real-world inequalities). Fairlearn, a Python toolkit, is used for building AI systems that are fairer. DP was chosen as a key fairness metric; it measures whether each group have similar proportions of positive outcomes. This promotes fairness by striving for all groups to be selected equally. EO was also measured, focusing on whether the error rates were equal for all groups. The model was mitigated using EG where training sample weights are adjusted based on DP. After applying mitigation techniques DP and EO were re-evaluated to assess the effectiveness of the mitigation attempt. The impact of the accuracy was analysed to understand any potential trade-offs. To conclude, ethical implications were addressed and recommendations for future works were suggested focusing on the importance of responsible AI development.

4. Data Collection and Analysis

The adult dataset will be explored, and an analysis will be presented.

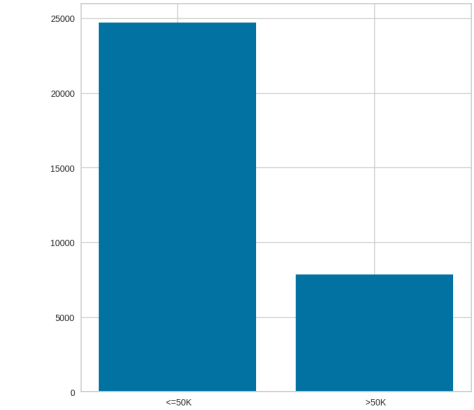


Figure 1: Distribution of Target

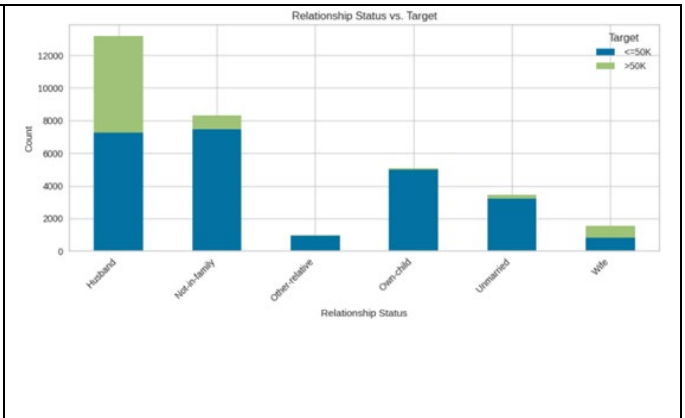


Figure 2: Relationship status vs target

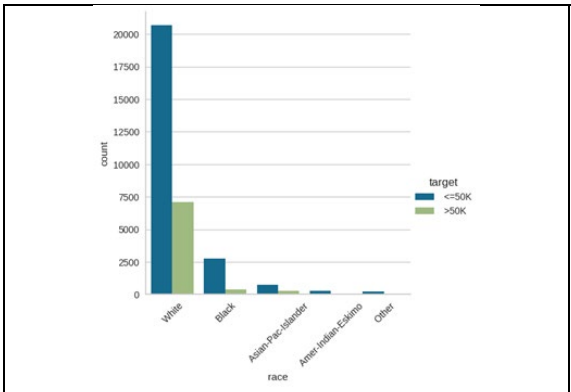


Figure 3: Distribution of race to target

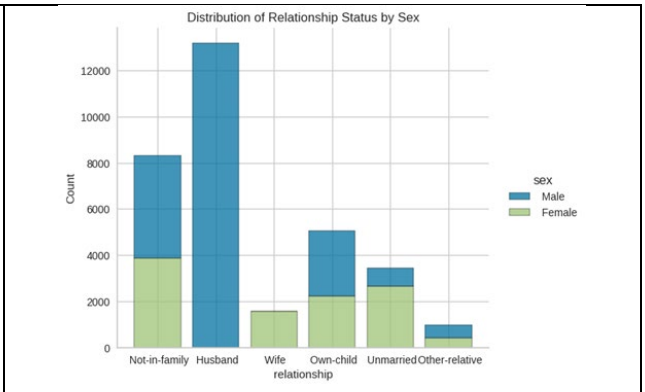


Figure 4: Distribution of relationship status by sex

4.1 Dataset Exploration

Explanatory Data Analysis(EDA) is performed on the adult dataset. EDA is the process of investigating the dataset to provide insights and better understanding. One of the goals of EDA is to uncover sensitive data attributes(sex, race, etc.), any correlations, and existing data quality issues in the dataset for potential bias. Python will be utilised to perform EDA on the adult dataset, including but not limited to pandas and others to visualise distributions and correlations. The goal is to use EDA to uncover potential biases such as unbalanced attributes and high correlations.

The dataset exhibits imbalances shown in Figures 3(racial) and 1(income). The over-representation of white individuals is an example of data bias. This imbalance can affect the model's ability to accurately predict income for those in the minority groups. Data bias is shown again in Figure 4, where Females are the underrepresented group. Fewer women make over 50k, potentially leading the model to favour males for higher income prediction.

Figure 4 reveals that 'Husband' perfectly correlates with 'Male' and vice versa making the relationship status attribute redundant. Husband is the most represented group, directly correlating with Figure 4 which shows men as the larger group. To address this, dropping the relationship attribute could be considered. Figure 5 highlights the unbalanced distributions within the relationship status attribute. The majority class is Husband again, while other categories lack high earners. Husband and wife classes have similar distributions, but the wife class is significantly underrepresented. This imbalance likely reflects historical bias and societal inequalities in earning potential between relationship groups(Norris et al., 2015).

Figure 5 highlights the correlation between all the features and the target variable. It highlights potential irrelevant features. Removing said features may be able to boost the accuracy score for the Gradient Boosting Classifier(GBC) Model. The features with the highest correlation are the education years, capital gain and the person's relationship status. The moderate positive correlation of 'education-num' and 'capital gain' suggests that those with higher education and capital gain are correlated with higher income.

The negative correlation for relationships needs more investigation. It is hard to tell which relationship status is negatively correlated as it has multiple classes. Figure 6 provides insights into the distribution in greater depth and Table 3 provides a label key for understanding. According to Figure 6, approximately 47% of wives are high earners, 45% of husbands are high earners and the other groups 7% are high earners. While wives are higher earners than husbands on average, the whole dataset will not reflect that as the class is outnumbered. As the husband group has the most individuals earning over 50k the ML model may inaccurately reflect that husbands make more than other relationship status.

Table 1: Relationship Status Value Label key

Relationship Status Value	Label
0	Husband
1	Not-in-family
2	Other-relative
3	Own-child
4	Unmarried
5	Wife

To summarise, this dataset is heavily unbalanced. If this dataset were used to train a model to detect whether an individual gets a loan by giving loans to a high earner it would show prejudice against those who are not a white married man.

5. Bias Mitigation Strategies

5.1 Bias Measurement

This section explores bias mitigation using Fairlearn, a Python toolkit created for building fairer AI systems. The metrics that will be used to measure the bias in the model are DP and EO. Unknown values were marked as '?' in the work class, occupation, and native country attributes were replaced with the mode to avoid data loss. Label encoding was used to relabel the categorical values to numerical ones for model interpretability. The data was split 80/20 for training and testing. Afterwards, SMOTE oversampling was performed on the training data to help mitigate the class imbalance of the target groups. GBCs are excellent models as they consistently rank

high for classification tasks. Boosting combines the predictions of multiple weak learners to create stronger and more accurate learners.

Figure 10 provides different accuracy metrics on the model classification outcome; the combined accuracy score of the model is 87%.

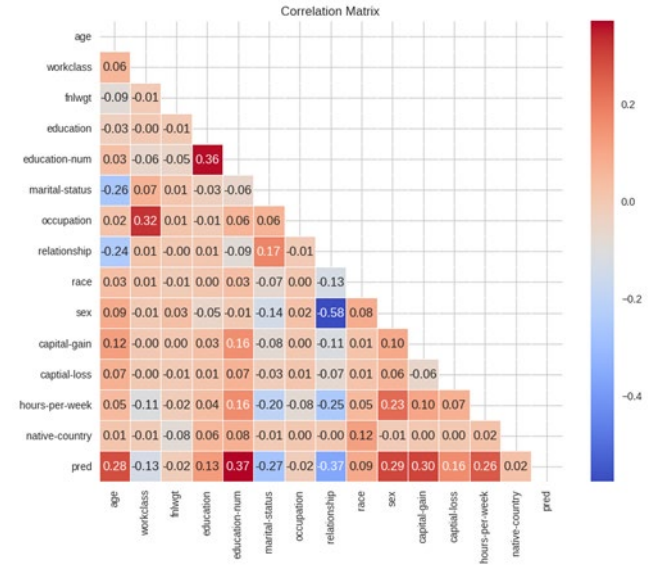


Figure 5: Correlation Matrix

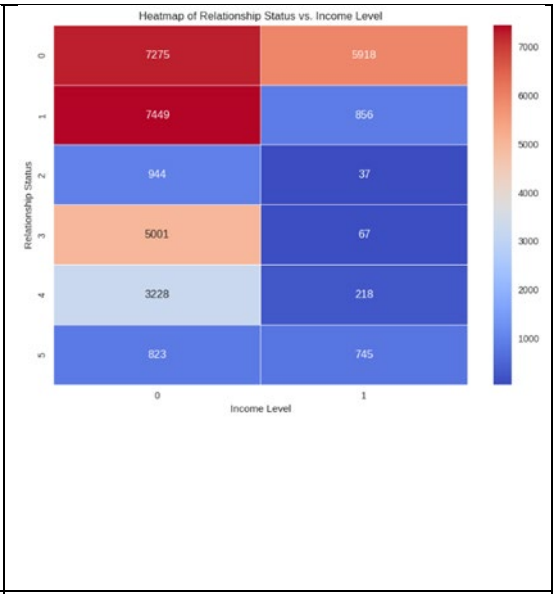


Figure 6: Heatmap of Relationship Status vs Income Level

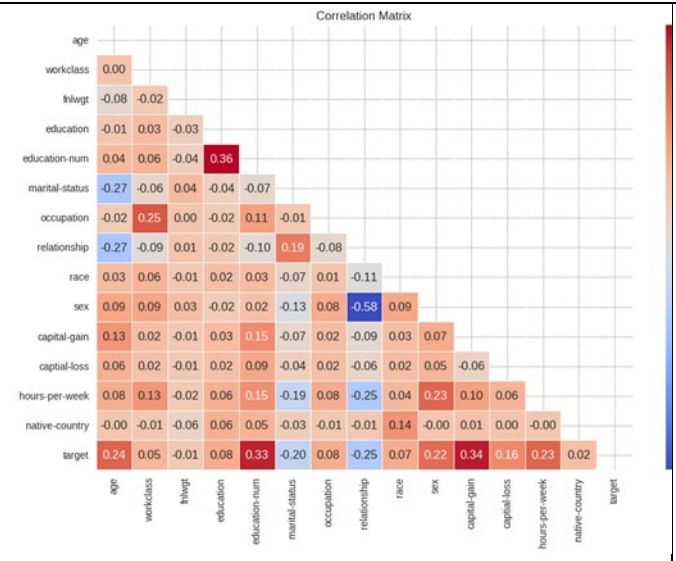


Figure 7: Correlation Matrix of Predicted Target

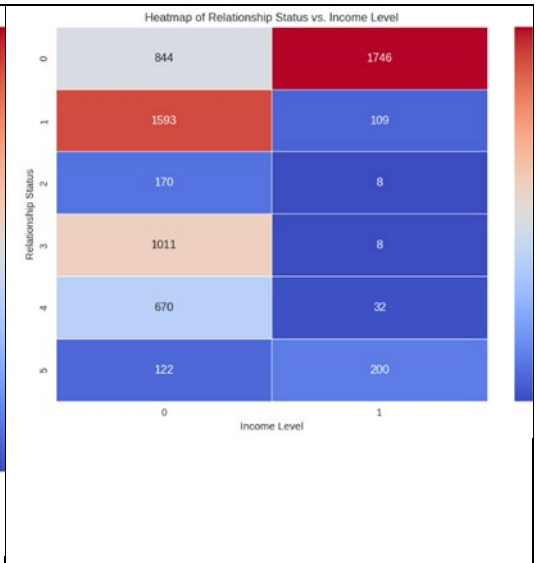
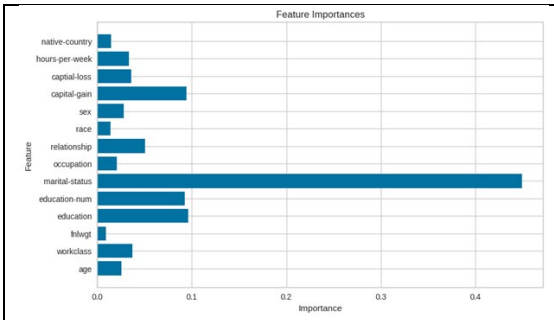


Figure 8: Heatmap of Relationship Status vs Predicted Income Level

Figure 7 is the correlation matrix of the test data and the targets the GBC model predicted. Compared with the correlation matrix on the data set before the ML model(Figure 5), the correlation matrix value for education-num, sex and relationship values have increased slightly; this is shown in Figure 7. This suggests that the algorithm slightly perpetuates the preexisting correlations. According to this diagram, education years positively correlate with higher earners while relationship negatively correlates. The distribution of high and low earners has significantly changed from Figure 8. In Figure 8 the percentage of women being labelled as higher earners is 62%, 67% for men and 6% for all other groups combined. While the original data showed that wives had a higher percentage of high earners, the husband class outnumbered the wife class. With this class imbalance, the ML

model is disproportionally trained on husband data, which in turn, causes the ML model to favour that class. Even with class balancing, the model has amplified the existing bias within the dataset. This is an example of algorithmic bias, where the ML model has learned and reinforced the biased patterns in the data. Feature importance refers to techniques that calculate a score for all the input features Shin(2023), explains that a larger score suggests more importance. According to Figure 9, marital status is the biggest deciding factor. Marital status has an importance value of 50% while all other attributes have values under 10% - this suggests very high importance; therefore, this feature likely has a powerful influence.



Class	Precision	Recall	F1-Score	Support
Under 50K	0.92	0.91	0.91	4942
Over 50k	0.72	0.74	0.73	1571

Figure 9: Feature Importances of the GBC Model

Figure 10: Accuracy score for GBC model

This could indicate that the model depends solely on one factor. From the EDA performed, the insights derived from that was that most individuals are married with the biggest group being husbands.

After identifying the biases, the next steps are to measure and quantify the biases to be able to compare values before and after mitigation processing. DP is a fairness metric that measures whether all groups have an equally positive outcome(Google for Developers, 2023). EO is another metric that will be utilised. Instead of measuring whether the model is predicting outcomes equally, EO measures the error rates for each class. The fairness metric calculates the true positive and negative rates, assessing whether the model is predicting outcomes equally well for all sensitive attributes(Google for Developers, 2024).

Figure 11 illustrates the DP difference between sexes, revealing a disparity of 0.22 in selection rates for high earners. This disparity is evidenced by Figure 14, which demonstrates that men(coded as 1) are more likely to be classified as high earners than women. It further supports how men are likelier to be falsely identified as a higher earner and women are more likely to falsely be identified as a lower earner through the error rates. This suggests that the ML model has learned the patterns that disproportionally benefit men and disadvantage women. Figure 12 provides the DP difference after mitigation using EG. The selection rate has been equalised between males and females resulting in the new DP score now being approximately 0.003. This shows that the disparities between the classes have been reduced, resulting in equal selection rates, and reducing the over-prediction of males.

```
0.22171531770618785
```

Figure 11: DP Difference of Sex before Mitigation

```
0.0028576905897244875
```

Figure 12: DP Difference of Sex after Mitigation

```
0.2967435862891895
```

Figure 13: Equalised Odds Difference on Sex attribute after mitigation

sex	accuracy	selection rate	false positive rate	false negative rate
0	0.886642	0.183443	0.105124	0.180258
1	0.835651	0.188967	0.034765	0.459641

Figure 14: Metrics for Sex attribute after Mitigation

5.2 Effectiveness of Mitigation Techniques

EG was a mitigation technique utilised to reduce the DP score. The weight change had a positive effect on the distribution of target classes. While the accuracy score was reduced(87% to 85%) the demographic parity scores were reduced(22% to 3%). This suggests that the classes now have an equal selection rate. Figure 14 reveals the complexities of bias mitigation. While a fairer DP has been achieved, the false negative rate of the male class has increased resulting in the EO gap widening. This suggests a trade-off between improving one fairness metric at the expense of another. Achieving parity often comes at the cost of higher error rates in specific groups. This highlights the challenge of different fairness considerations. Additionally, the accuracy score has been reduced when implementing mitigation methods. This also illustrates the trade-off between fairness and accuracy. The next section explores the ethical dilemmas of biased algorithms, their negative consequences, and the role of AI developers in mitigating these issues.

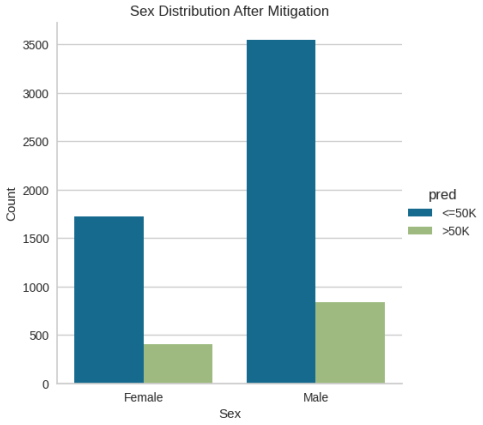


Figure 15: Sex distribution after mitigation

6. Ethical Considerations

This section will explore the ethical dilemmas of biases within AI and the repercussions., this will highlight the negative effects of biased AI systems to reiterate the importance of mitigation. After exploring the ethical dilemmas, implications, and social and legal consequences of those biases; strategies for responsible AI development will be discussed.

6.1 Ethical Dilemmas

Ethical dilemmas arise from biased AI systems, affecting various sectors like health, finance, and system development processes Ludovica Ilari et al.(2023). Khare(2023) explains that biased AI systems can have severe consequences for individuals and society. For example, if an AI scoring system is biased it can perpetuate financial inequalities; making it difficult for individual groups in marginalised communities to access credit. The adult dataset poses an ethical challenge, the data reflects historical and socioeconomic bias which raises questions about using datasets that perpetuate biases to train AI models. With efforts to mitigate biases through over-sampling and EG, disparities were still observed.

6.2 Ethical and Societal Implications of Biased AI

As AI becomes integrated into our lives, the potential for technologies to shape cultural norms and social structures becomes more significant(Ferrara, 2023). This could lead to new forms of discrimination based on skin colour or physical appearance. It is even possible for humans to inherit AI bias, according to a study conducted by Vicente and Matute(2023) where participants completed a medical-themed classification task with the help of a biased AI system. When participants were asked to complete the task again without assistance, they mimicked the same mistakes that biased AI systems had made during the previous phase. This suggests that biased AI could potentially shape future societal norms.

Additionally, bias amplification is another issue that causes societal and social implications. When initial models are biased, they produce biased outputs which are later used to train newer models resulting in the newer models learning more biases. To determine the effects, Bohdal et al.(2023) explained that if a job hiring decision AI is based on previous hiring decisions, bias in the subgroup can reinforce more bias in the future. This could

lead to worse jobs and worse access to finance in that subgroup. Over time the subgroup would be socially and economically marginalised, widening existing inequalities.

6.3 Responsibilities of AI Developers

Ethical considerations, transparency and explainability in models must be prioritised. Developers should address biases by fostering diversity in their teams and conducting testing in diverse societal groups(Sanderson et al., 2023). Engaging with diverse stakeholders, including marginalised groups, throughout AI development can ensure these systems meet everyone's needs(Petar Radanliev et al., 2024). This ensures that the training data is diverse, while actively identifying and addressing any potential biases present in the data. Developers need to be able to apply mitigation techniques, to reduce any discriminatory outcomes. Doris and Potter(2024) suggest continuous monitoring, this allows developers to proactively identify any new biases. By following these principles developers can build more trust into these systems, ensuring that AI serves as a tool for good rather than harm.

Overall, this section explored the ethics surrounding biases within AI, highlighting the potential of societal harm due to unfair systems. The adult dataset served as a case study to showcase the presence of biases and demonstrate the challenges when mitigating biases and their trade-offs with accuracy scores and fairness metrics.

7. Conclusions and Future Work

This section summarises the key findings of this research and offers practical guidelines for future AI developers in promoting fair and ethical AI practices.

7.1 Summary of Findings

This paper investigated the potential biases in ML models trained on the Adult Census Income dataset, focusing on the relationship between income prediction and sensitive attributes such as marital status. Mitigation techniques were applied, and the disparities were reduced. While the dataset is valuable for income prediction, it contains sources of bias that could negatively impact model fairness. The bias in the ML model manifested from the over-representation of the male group resulting in the ML disproportionality predicting men as high earners than women. If this dataset were solely to be used for training a credit card application model; women would have a harder time getting credit, perpetuating financial inequalities. These biases were addressed through preprocessing, feature engineering, and fairness awareness metrics. This was to strive for a model that would produce equal outcomes for all individuals regardless of their demographic characteristics.

While a fairer model was achieved it had its trade-offs in accuracy and equal odds. The false negative rate for men increased suggesting that men were more frequently misclassified as low earners compared to women. While this was the case, the number of men being predicted as higher earners was still higher than women. This poses an ethical dilemma, even though the goal was to reduce bias, the mitigation strategy has introduced a new form of unfairness. This poses a question of whether DP should be prioritised at the expense of equalised odds. In this case for this dataset, DP should be prioritised as EO could not be achieved.

7.2 Implications of Findings in the AI Field

This section will provide a brief overview of how the findings in this dissertation contribute to the field of AI.

This paper has demonstrated an understanding of how bias manifests within ML models, often originating from the data itself. It contributes to data bias awareness, a critical issue as data bias is the biggest source of bias in ML models. The findings support existing theories on how biases are present in data and how they are further perpetuated, often through algorithmic amplification. This study's findings challenge the assumption that bias is solely created by the ML model and reiterate that bias can be introduced at several stages of an algorithm's development. This paper offers new insights into whether all mitigation steps should be taken, even if the trade-off is a worse-performing model, revealing an ethical dilemma on whether it is worth under-predicting certain groups in the name of fairness. Overall, by raising awareness of these issues and exploring potential solutions, this research contributes to the ongoing efforts to develop fairer algorithms.

7.3 Recommendations for Future Works

Future research should focus on developing stronger fairness metrics and exploring more effective mitigation techniques that address bias at every stage in the ML pipeline. Synthetic data generation should be considered to lessen the impact of overrepresented groups where more data collection is not possible. For future

developers, they should focus on potentially developing new algorithms to specifically address fairness issues in models, users need to understand, trust, and manage AI results.

Encouraging collaboration between AI researchers and experts in social science, ethics, and law to ensure that AI development is guided by diverse perspectives with an understanding of societal impact is essential.

7.4 Guidelines

AI systems should be lawful, ethical and robust as discussed in the European Commission(2019).

References

Akselrod, O.(2021). How Artificial Intelligence Can Deepen Racial and Economic Inequities | ACLU. [online] American Civil Liberties Union. Available at: https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities .

Becker, B. and Kohavi, R.(1996). UCI Machine Learning Repository. [online] Uci.edu. Available at: https://archive.ics.uci.edu/dataset/2/adult.

Bohdal, O., Hospedales, T., Philip and Barez, F.(2023). Fairness in AI and Its Long-Term Implications on Society. [online] arXiv.org. Available at: https://arxiv.org/abs/2304.09826#:~:text=AI%20fairness%20focuses%20on%20mitigating,act%20as%20a%20social%20stressor.

Chopra, D. and Khurana, R.(2023). Bias and Fairness in Ml. BENTHAM SCIENCE PUBLISHERS eBooks, [online] pp.116–122. doi:https://doi.org/10.2174/9789815124422123010012.

Doris, L. and Potter, K.(2024). Continuous Monitoring and Improvement: Implement continuous monitoring of AI models to detect and correct... [online] ResearchGate. Available at: https://www.researchgate.net/publication/378342382_Continuous_Monitoring_and_Improvement_Implement_continuous_monitoring_of_AI_models_to_detect_and_correct_issues_in_real-time#:~:text=By%20continuously%20monitoring%20AI%20models,continued%20effectiveness%20of%20deployed%20models.

European Commission(2019). HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINES FOR TRUSTWORTHY AI. [online] Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

Ferrara, E.(2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci, [online] 6(1), pp.3–3. doi:https://doi.org/10.3390/sci6010003.

González-Sendino, R., Serrano, E. and Bajo, J.(2024). Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. Future Generation Computer Systems, [online] 155, pp.384–401. doi:https://doi.org/10.1016/j.future.2024.02.023.

Google Cloud.(2024). AI vs. Machine Learning: How Do They Differ? | Google Cloud. [online] Available at: https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning .

Google for Developers.(2023). Machine Learning Glossary: Fairness. [online] Available at: https://developers.google.com/machine-learning/glossary/fairness#:~:text=demographic%20parity&text=A%20fairness%20metric%20that%20is,on%20a%20given%20sensitive%20attribute.

OV.UK(2023). New innovation challenge launched to tackle bias in AI systems. [online] GOV.UK. Available at: https://www.gov.uk/government/news/new-innovation-challenge-launched-to-tackle-bias-in-ai-systems .

Ludovica Ilari, Giulia Rafaiani, Baldi, M. and Benedetta Giovanola(2023). Ethical Biases in Machine Learning-based Filtering of Internet Communications. [online] doi:https://doi.org/10.1109/ethics57328.2023.10154975.

Markou, G., Bakas, N. and Megan, A.(2023). Use of AI and ML Algorithms in Developing Closed-Form Formulae for Structural Engineering Design. Advances in civil and industrial engineering book series, [online] pp.73–105. doi:https://doi.org/10.4018/978-1-6684-5643-9.ch004.

Masha Fedzechkina, Lucy Hall Hartley and Roberts, G.(2022). Social biases can lead to less communicatively efficient languages. [online] 30(3-4), pp.230–255. doi:https://doi.org/10.1080/10489223.2022.2057229.

Nazer, L., Razan Zatarah, Waldrip, S., Janny, Moukheiber, M., Khanna, A.K., Hicklen, R.S., Lama Moukheiber, Moukheiber, D., Ma, H. and Mathur, P.(2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS digital health, [online] 2(6), pp.e0000278–e0000278. doi:https://doi.org/10.1371/journal.pdig.0000278.

Park, Y. and Hu, J.(2023). Bias in Artificial Intelligence. PubMed, [online] 18(3), pp.394–396. doi:https://doi.org/10.2215/cjn.0000000000000078.

Parliament.uk.(2023). Artificial Intelligence(Regulation) Bill [HL] - Parliamentary Bills - UK Parliament. [online] Available at: https://bills.parliament.uk/bills/3519.

Petar Radanliev, Santos, O., Alistair Brandon-Jones and Joinson, A.(2024). Ethics and responsible AI deployment. Frontiers in artificial intelligence, [online] 7. doi:https://doi.org/10.3389/frai.2024.1377011.

Raji, I.D. and Buolamwini, J.(2019). Actionable Auditing. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. doi:https://doi.org/10.1145/3306618.3314244.

Shin, D. and Shin, E.Y.(2023). Data's Impact on Algorithmic Bias. IEEE Computer, [online] 56(6), pp.90–94. doi:https://doi.org/10.1109/mc.2023.3262909.

Shin, T.(2023). Understanding Feature Importance in Machine Learning. [online] Built In. Available at: https://builtin.com/data-science/feature-importance .

The Telegraph(2023). Rishi Sunak Warns of Risks Posed by AI in Tech Summit. YouTube. Available at: https://www.youtube.com/watch?v=262YJQWE7Ec.

Vicente, L. and Matute, H.(2023). Humans inherit artificial intelligence biases. Scientific reports, [online] 13(1). doi:https://doi.org/10.1038/s41598-023-42384-8.

Wamburu, J., Girmaw Abebe Tadesse, Cintas, C., Adebayo Oshingbesan, Akumu, T. and Speakman, S.(2022). Systematic Discovery of Bias in Data. 2022 IEEE International Conference on Big Data(Big Data). [online] doi:https://doi.org/10.1109/bigdata55660.2022.10020781.

Yang, J., Clifton, L., Dung, N.T., Phong, N.T., Yen, L.M., Thy, D.B.X., Soltan, A.A.S., Thwaites, L. and Clifton, D.A.(2024). Mitigating machine learning bias between high income and low–middle income countries for enhanced model fairness and generalizability. Scientific Reports, [online] 14(1). doi:https://doi.org/10.1038/s41598-024-64210-5.

Zavaleta-Sánchez, E., Domínguez-Sánchez, G., Cecilia-Irene Loeza-Mejía and Sánchez-DelaCruz, E.(2024). Comparative Study of KDD and CRISP-DM Methodologies for Phishing Identification. Lecture notes in networks and systems, [online] pp.317–330. doi:https://doi.org/10.1007/978-981-97-3559-4_25.

Designing an Artificial Intelligence Maturity Model for Human Resources (HR-AIMM)

Sascha Armutat, Malte Wattenberg and Nina Mauritz
Bielefeld University of Applied Sciences and Arts, Germany

sascha.armutat@hsbi.de
malte.wattenberg@hsbi.de
nina.mauritz@hsbi.de

Abstract: Artificial intelligence (AI) has the potential to change the world of work radically. Wherever information processing is involved, AI can be integrated into processes with added value. From the perspective of Human Resource (HR) management, this implies three things: first, business models and performance processes in the company will undergo change; second, employee requirements will change; and third, HR processes will change. While the literature describes various AI maturity models, there has been no dedicated consideration of HR management. This article, therefore, aims to identify relevant influencing factors for an AI-orientated approach to HR management and to describe these in more detail using maturity levels in a Human Resources Artificial Intelligence Maturity Model (HR-AIMM). The resulting HR-AIMM consists of eleven dimensions. These include anchoring the AI topic in the corporate and HR strategy, its use in selected HR processes, considering ethical, data-related, and infrastructural principles, and organisational, cultural, and competence-related anchoring. The characteristics of these factors enable the identification of four maturity levels for using AI in HR management: from a curious start to the level of holistic integration. Our framework supports researchers and companies in understanding and evaluating the factors influencing the professional application of AI in HR management.

Keywords: Maturity model, Artificial intelligence, Human resources, HR-Management, AI capability, AI readiness

1. Changing the World of Work Through Artificial Intelligence

Artificial intelligence (AI) has the potential to radically change the world of work (Franken and Wattenberg, 2019). The overwhelming interest in this technology can be seen, among other aspects, in numerous previous attempts to define it (see Palos-Sánchez et al, 2022). Haenlein and Kaplan (2019, p. 1) view AI as *"a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation"*.

Since the race to improve generative artificial intelligence became public with ChatGPT, the extent of the potential changes has become increasingly apparent. The economic opportunities are enormous: McKinsey calculates that generative AI, in particular, has a value potential in the range of USD 6.1 - 7.9 trillion – a result that can be attributed to new use cases on the one hand and to productivity increases among employees on the other (Chui et al, 2023).

Accordingly, there is a high interest in implementing AI in organisations.

At the same time, there are massive implications for work processes and workforce structures: Generative AI will replace a proportion of working hours in all jobs. Hazan et al (2024) have calculated that, on average, generative AI can replace 27% of the hours employees work. The more administrative and routine-based a job is, the more likely AI will substitute parts of the activity. Occupations characterised by standardised information processing are also at risk (Eloundou et al, 2023). Overall, companies assume that just as many jobs will be created through AI as will be lost (Hays, 2024).

Human Resource (HR) management faces two challenges in this change situation: Firstly, it must actively support the integration of AI into company processes and structural and skills-related changes within the company – as a change agent, organisational and HR developer, and strategic partner to management. Secondly, it must adapt its processes, structures, tools, and competencies considering the possibilities of AI-supported HR work.

It is precisely because HR management has a role model and pioneering function due to its role in supporting change that it is necessary to take a closer look at the company's own AI introduction and implementation processes.

This is where maturity models (MM) step in, enabling companies to assess and monitor their status quo in a specific domain (Alsheiabni et al, 2019; Wendler, 2012). The main idea of an MM is *"that it describes in a few phrases, the typical behaviour exhibited by a firm at a number of levels of maturity, for each of several aspects"* (Fraser et al, 2002, p. 244).

Despite considerable interest from scholars in the development of AI-related maturity models, there needs to be a greater understanding of the dimensions and maturity levels relevant to the specific domain of HR.

For this reason, this article aims to close this gap by operationalising relevant dimensions for using AI in HR and transferring them into an MM (HR-AIMM).

Therefore, the aim is to develop and transfer an MM with the dimensions of influencing factors that need to be considered in the professional implementation and application of AI in HR management.

In close accordance with previous research on MM (e.g. Fukas et al, 2021), the underlying questions of this study are:

RQ1: Which dimensions and components represent the influencing factors for the HR domain concerning the implementation of AI technologies?

RQ2: How can the dimensions and components be described using maturity levels and mapped in a modular, multidimensional maturity model?

The remainder of the paper is structured as follows: First, Chapter 2 addresses the state of research on AI maturity models. This is followed by a description of the method in Chapter 3 and the presentation of a model in Chapter 4. The paper concludes with a discussion in Chapter 5 and a conclusion in Chapter 6.

2. Related Work

The related work has been organised into two main sections. The first summarises HR and AI literature, and the second focuses on the MM approach to AI.

2.1 Human Resource Management

For some time now, companies have recognised the relevance and inevitability of AI in HR to adapt to rapidly changing conditions and withstand competitive pressure.

Boselie et al (2021, p. 484) emphasise that "*HRM involves management decisions related to policies and practices that together shape the employment relationship and are aimed at achieving certain goals"*. Therefore, while Boselie et al (2021) focus on management decisions and their influence on the employment relationship and thus emphasise the strategic orientation of HRM to achieve organisational goals. HRM is often operationalised as a combination of different HRM practices that shape employee relations within and outside the organisation (Boselie et al, 2021).

The application of AI in HRM is an emerging field with continuous growth and a promising outlook for the future (Palos-Sánchez et al, 2022). AI is reshaping processes in almost all critical areas of HRM and is increasingly being integrated into various operational HR processes (Tambe et al, 2019).

Current areas of HRM AI applications can be found throughout the entire employee life cycle, starting with recruitment, selecting suitable candidates, onboarding, performance management, training and development, and retention (Kaushal et al, 2023). Various AI-based technologies, such as automated systems, personal assistants or chatbots, can be used. A study of HR professionals confirms that 89% already use AI tools in their HR departments. The most common use case is generative AI systems such as ChatGPT in recruitment and hiring processes (59%). Overall, communication with human contact persons is being severely restricted (Greenhouse, 2023).

Advanced digital technologies such as AI create added value and cost efficiency, for example, by facilitating management decision-making through an expanded range of knowledge (Kumar et al, 2022). Although AI can thus optimise all phases of HRM, there are numerous challenges that HRM must face when implementing AI (Palos-Sánchez et al, 2022; Tambe et al, 2019). These include managing the overall impact of AI on employees (A. Malik, 2023), leadership resistance (Frick et al, 2021), human rights and ethical challenges regarding data privacy and discrimination (Stahl et al, 2023).

2.2 Maturity Models

The research literature uses the terms "*maturity*" and "*readiness*" as starting points for developing models.

Maturity generally refers to a system's progress towards a target state, defined by a series of successive stages.

AI readiness, in turn, refers to "*the preparedness of organisations to implement change involving applications and technology related to AI*" (Alsheibani et al, 2018, p. 3).

Depending on the authors, "*readiness"* is seen as preceding, synonymous, embedded or context-dependent to "*maturity"* (Reichl and Gruenbichler, 2023).

Although there is great interest in maturity models (MM) in the research landscape, only a few authors provide a definition. Pullen (2007, p. 9) defines MM as a "*structured collection of elements that describe the characteristics of effective processes at different stages of development [and] also suggests points of demarcation between stages and methods of transitioning from one stage to another*".

An MM thus describes the typical development paths of an object class by mapping development stages, from the starting point at the lowest level to full maturity at the highest level in the domain under consideration (Becker et al, 2009). As a result, MMs are helpful tools for describing a specific domain's status, potential and requirements (Wendler, 2012). They allow organisations to monitor the step-by-step development within an implementation process, leverage the capabilities of a particular domain and increase its strategic potential (Alsheiabni et al, 2019; Bruin et al, 2005). They serve as a starting point for future growth, identify necessary steps and potential transition challenges, and help prioritise. Through generally accepted growth stages, they enable the definition of progress and the measurement of improvements compared to other organisations (Pullen, 2007). Therefore, MMs are a strategic tool for continuous comparison and roadmap development (Fukas et al, 2021).

Furthermore, MMs can be classified according to various characteristics. Bruin et al (2005) distinguish between three application-related purposes of MMs: 1) Descriptive: evaluation of the current state, 2) Prescriptive: normative models that offer specific recommendations for action and guidelines for growth, 3) Comparative: models with which companies can be compared internally and externally. In addition, the scope of MMs can be general or domain-specific and can differ according to the level of analysis (company, department, people, projects, system, process, object) (Sadiq et al, 2021). The model itself can be structured cyclically or iteratively, whereby the degree of maturity can be measured qualitatively or quantitatively as well as discretely or continuously (Kohlegger et al, 2009). However, according to Sadiq et al (2021), standardised terminology for modelling (e.g. dimensions, constructs, elements, indicators) has yet to be established despite numerous MMs.

Even though MMs are widely represented in the research literature, this number decreases considerably with an additional increase in specificity and a reference to AI.

Lichtenthaler's (2020) framework offers an initial approach in which the author presents five maturity levels of AI management. In addition, Saari et al (2019) developed an AIMM to evaluate the maturity level of different departments in the company. The conceptual AIMMs by Alsheiabni et al (2019), Limat (2022) and Jaaksi (2018) are located at the organisational level. The same applies to Schuster et al (2021), focussing on small and medium-sized enterprises. Holmström (2022) and Jöhnk et al (2021) form a framework with indicators that characterise the readiness of companies, although Jöhnk et al (2021) do not define maturity levels. The "*3-horizon model"* by Kreutzer and Sirrenberg (2020) is to be understood as an AI maturity map, which also considers fields of application and should serve as a starting point for a company's own AI journey. Other AIMMs relate to the domains of marketing (Gentsch, 2019), innovation management (Yams et al, 2020), logistics (Ellefsen et al, 2019), manufacturing (Sonntag et al, 2024) and auditing (Fukas et al, 2021).

In addition, developing maturity models is a highly competitive market to which various organisations such as market analysts, consulting companies and software providers are committed. AIMMs from these sources include ElementAI (2020), appliedAI (2021), Accenture (2022), MITRE (2022), DFKI (2022) and Deloitte (2024). Despite lacking scientific documentation, the advantages of using AIMMs from these organisations may include greater practical suitability and (online) self-assessments.

Despite presented AIMMs cover several different domains, a significant research gap remains: To the authors' best knowledge, no study currently addresses the domain of HR, though this is crucial for a holistic understanding of a company's maturity.

3. Method

The procedure of the present study is based on the steps widely accepted and used in research for the development of a MM by Becker et al (2009), which is based on the Design Science Research Approach. In

addition, the recommendations of Bruin et al (2005) are included in the development (fig. 1). The approach is therefore presented using a linear logic with strong interdependencies.

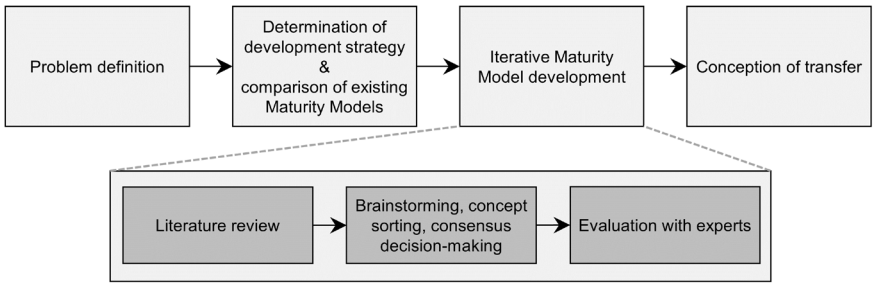


Figure 1: Steps of MM development based on Becker et al (2009) and Bruin et al (2005). Source: Author's own work

3.1 Problem Definition

According to Becker et al (2009) the initial stage of the process involves defining the problem, which includes identifying the targeted domain and target group, discussing the relevance of the problem and the anticipated benefits, and determining the conditions for applying the model. With this in mind, HR was defined as the target domain. The intended MM aims at HR managers and company executives. Next, it is necessary to outline the research objectives and formulate key questions. Finally, there is the need to concentrate on refining its functionality, effectiveness, and clarity, thereby validating the conditions necessary for its application.

3.2 Determination of Development Strategy and Comparison of Existing Maturity Models

Since, to the best of the authors' knowledge, no MM for AI could be identified in HR at the time of the study, we follow the procedure for developing an entirely new MM according to Becker et al (2009).

3.3 Iterative Maturity Model Development

This process step's objective was the fundamental development of the model, including the definition of the modelling approach, the elements, and the validation.

Regarding the modelling approach, a prescriptive model with a multidimensional structure was chosen due to its high relevance, transferability into practice, and the expected complexity. As there was no HR-AIMM and, therefore, no preliminary basis for the development of the model, a top-down approach was chosen in which the maturity levels are developed first and subsequently the items of assessment (Bruin et al, 2005).

For the development of the elements, this study is based on the sub-processes of Rigamonti et al (2024), including literature research, knowledge-generating techniques, and validation.

In the first sub-process, the existing literature must be analysed (Becker et al, 2009). In a first step, existing systematic literature reviews (SLR) of AIMMs were researched, and the works of Sadiq et al (2021) and Reichl and Gruenbichler (2023) were identified. Cited studies included in both SLRs were combined and analysed for cardinality: $|A \cup B| = |A| + |B| - |A \cap B|$ with A=SLR(A) and B=SLR(B). Given that the vast majority of literature on the development of AIMMs also contains SLRs for existing MMs, further AIMMs were identified following a forward and snowball search. In the second step, we conducted our research using a slightly formalised narrative review using the search term (*"artificial intelligence maturity model" AND "human resources")* and its German translation on Google Scholar and Google. The reason for this approach is that previous SLRs have lacked the results of the efforts of numerous practice-oriented organisations and institutes due to their methodologically rigorous selection of research literature. According to this paper's authors, however, including these sources with a practical perspective is decisive for developing a transferable MM. In the third step, identified studies were compared in tabular form based on the dimensions and maturity levels used. While a complete review of AIMMs is beyond the scope of this article, Chapter 2.2 already outlined several examples. Finally, a synopsis was created.

Various knowledge-generating techniques were used as the following sub-process (McGraw, 1989). Following Rigamonti et al (2024), several brainstorming sessions with the research team were held to identify dimensions and maturity levels for a preliminary MM. Concept-sorting, a technique in which participants sort categories according to their subjective similarity, was used to structure dimensions, subdivide them into

further components, and assign maturity levels. Finally, consensus decision-making was used to achieve agreement among all participants, favouring solutions supported by all participants.

The final sub-process relates to evaluating the formulated maturity levels (Becker et al, 2009) to ensure the validity of the content. To this end, a qualitative research study was conducted with six HR and training experts from different industries. The experts were selected based on their professional experience, personnel and organisational development knowledge, and digitalisation topics. They received the maturity levels for written assessment and were evaluated according to expression, comprehensibility, clarity, appropriateness, and differentiation criteria. They were also able to leave detailed comments. The written feedback was analysed using a qualitative content analysis. Moderated discussions were held via Zoom, during which the ratings and comments were discussed and documented. The analysis of this documentation revealed strengths and weaknesses of the maturity levels, particularly concerning complexity, technical language and unclear terms. Suggestions for clarification were developed. Feedback on the appropriateness of the maturity levels about the practical requirements was also included.

3.4 Conception of Transfer

The final step is to define the transfer medium (Becker et al, 2009). For this purpose, an interactive website was created that enables companies to determine their level of maturity of AI implementation in HR by answering a questionnaire. In addition, presenting results in a radar chart allows benchmarking with comparable companies (Armutat et al, 2024).

4. Structure of the HR AI Maturity Model

The HR-AIMM consists of eleven dimensions with two items each (see Table 1) and four maturity levels (cf. 4.1).

4.1 Maturity Levels

The following maturity levels were defined:

- Curious start: Companies have not yet undertaken any activities relating to the integration of AI in HR management.
- Learning experimentation: Companies are experimenting with AI-supported systems for selected groups of people or in individual areas of HR management.
- Project-related implementation: Companies already have initial experience with the task- or project-related introduction of AI-supported systems in HR management.
- Holistic integration: Working with AI systems is established and anchored in various areas of HR management within the company.

4.2 Dimensions

Table 1 below presents the dimensions and components:

Table 1: Description of dimensions. Source: Author's own work. Note: Core dimensions were summarised due to identical items and thus do not accumulate to 11

Dimension (d)	Item (i)	Differentiation by maturity level (m1-m4)
People and culture	Change Management	change management is (1) no topic yet, (2) has been considered, (3) planned or (4) established
	Participation	employees are (1) not informed, (2) informed, (3) partially involved or (4) involved in collaborative, cross-functional teams for further development
AI in corporate strategy	Role of AI	(1) AI plays no role, (2) is being experimented with, or whether (3) strategic projects have been initiated or (4) established
	Attitude of the management	the company management (1) has not positioned itself, (2) is open to experimentation, (3) demands AI measures or (4) has firmly established them
Data Management	Data collection and data utilisation	data collection and data use have (1) not yet been addressed, whether the prerequisites (2) are being worked on or whether a DMS (3) has been implemented and (4) automated
	General conditions	regulations are (1) not in place, (2) planned, (3) occasional or (4) extensive

Ethics	Conception of AI guidelines	guidelines are (1) not in place, (2) planned, (3) introduced or (4) established with compliance processes
	Integration of AI guidelines	measures for compliance with AI guidelines are (1)not in place, (2) planned, (3) introduced or (4) established
Infrastructure	Use of AI technologies	AI technologies are (1) not in use, (2) researched, (3) under development or (4) implemented
	Required hardware	the necessary hardware (1) is known, (2) is being procured and (3) has been partially or (4) fully implemented
AI in HR strategy	Relevance	AI is (1) no part of the HR strategy, (2) experiments with AI in individual HR tasks are strategically anchored, (3) strategic AI projects for selected HR processes are anchored, (4) AI-based HR processes are a central component
	Competence status of employees	employees have (1) no, (2) partial or (3) full competencies and (4) develop these further
Competence development	Competence profiles	necessary changes in skills are (1) unknown or (2) analysed, and whether target profiles have been (3) partially or (4) fully derived
	Development of AI-related competencies	development measures are (1) available, (2) planned, (3) partially or (4) entirely realised
Organisation	Processual anchorage	processes are (1) undefined, (2) partially, (3) coordinated or (4) fully integrated
	Personal responsibility	a responsibility is (1) non-existent, (2) project-related, (3) temporary or (4) permanent
AI in HR recruitment, HR deployment, HR development	Experience	there is (1) no, (2) experimental, (3) task-related or (4) optimised experience
	Attitude	HR stakeholders are (1) negative or (2) open to AI and (3) share experiences or (4) are experienced in its use

4.3 Visualisation and Calculation

The MM can be visualised as a box model in which some dimensions can be structured based on our findings in such a way that they serve as a framework for other dimensions (Figure 2).

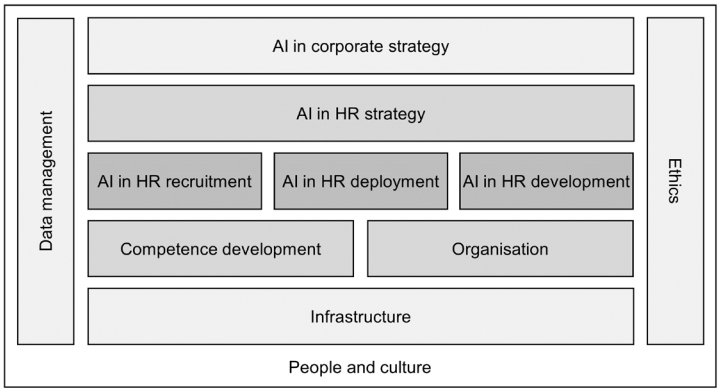


Figure 2: Multidimensional maturity model with framing and independent dimensions. Source: Author's own work

However, there are only slight interdependencies between the dimensions, which means that the level of development in one dimension does not necessarily influence or condition the maturity in another. This allows independent assessment and further development of each dimension, which promotes flexibility and adaptability in implementing and applying the model. Nevertheless, a distinction between supporting dimensions in three layers and exclusive HR core dimensions can be made. The maturity level of all dimensions is calculated using the function:

$$M_{model} = \sum_{d=1}^{11} \sum_{i=1}^{2} m_{(d,i)}$$

Whereby:

d: Dimension index with d ∈ {1,2,…,11}.

i: Item index within one dimension, with i ∈ {1,2}.

m: Maturity level of an item, whereas $m_{d,i}$ represents the maturity level of item i in dimension d with $m_{d,i} \in \{1,2,3,4\}$.

The average maturity level is therefore calculated as follows:

$$\bar{M} = \frac{M_{model}}{D \times I}$$

5. Discussion

5.1 Practical Contribution

The maturity model has a wide range of possible applications in companies. Understood as an operationalised maturity level assessment, it helps decision-makers in the company to analytically determine the status quo that needs to be considered in the implementation process. At the action level, activities necessary to create the conditions for the successful introduction of AI in HR management can be derived - be it data-related challenges or, for example, the development of the competencies of the relevant stakeholders. With this help, the model can support project planning and help to identify appropriate work packages. Secondly, the model helps to develop a perspective for AI integration in the HR management strategy: Relevant strategic target areas come into view based on a holistic strategic approach that gives equal weight to structures, competencies and cultural framework conditions. Thirdly, it could be a continuous periodic indicator of how HR management is developing in the direction of AI. This allows culturally sensitive and transparent introduction processes to be planned and implemented. Involving the workforce promotes the acceptance of AI applications and the maturity check can be used for introduction processes beyond HR management.

5.2 Theoretical Contribution

To the authors' knowledge, the proposed MM is the first HR-AIMM described in the scientific literature. This closes a remarkable research gap and offers researchers numerous reference points.

5.3 Limitations

Our study is subject to certain limitations frequently encountered in the development of MMs: It focuses on the actual state and the potential, while specific recommendations for action to achieve the following maturity levels can only be derived from the description of the items. In addition, the interdependencies of the dimensions shown in the box model were only validated to a limited extent and can, therefore, only be regarded as initial indicators. As interdependencies become more critical with the increase in complex organisational skills, this is a particular opportunity for further follow-up research.

6. Conclusion

The article shows which influencing dimensions need to be considered in the professional implementation and application of AI in HR management so that the use of AI systems increases the success of the HR function. The systematic analysis, including MMs on the market and the evaluation, ensures that the maturity levels are theoretically sound, practically relevant, and understandable. The framework can be mapped in a practical questionnaire that interested individuals or HR managers can complete. In this way, it helps companies to systematically analyse the status quo of AI requirements and use in HR management and thus provides orientation for the strategic development of technically supported HR work. ollow-up research may also feel particularly encouraged to address the interdependencies between dimensions, as these become increasingly important with the growth of complex organisational capabilities.

Acknowledgements

The results of this paper are based on the research project Kompetenzzentrum Arbeitswelt.Plus. This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within “The Future of Value Creation – Research on Production, Services and Work” program and managed by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the content of this publication.

The proposed maturity model can be assessed via an online self-assessment at https://hr-aimm.com.

References

Accenture. (2022) “The art of AI maturity. Advancing from practice to performance”, [online], https://www.accenture.com/content/dam/accenture/final/a-com-migration/manual/r3/pdf/pdf-4/Accenture-Art-of-AI-Maturity-Report.pdf.

Alsheiabni, S., Cheung, Y. and Messom, C. (2019) “Towards An Artificial Intelligence Maturity Model: From Science Fiction To Business Facts” *PACIS Proceedings(46)*, https://aisel.aisnet.org/pacis2019/46.

Alsheibani, S., Cheung, Y. and Messom, C. (2018) “Artificial Intelligence Adoption: AI-readiness at Firm-Level” *PACIS Proceedings(37)*, https://aisel.aisnet.org/pacis2018/37.

appliedAI. (2021) “AI Maturity Assessment”, [online], https://www.appliedai.de/en/solutions-services/strategy-transformation/ai-maturity-tool.

Armutat, S., Wattenberg, M. and Mauritz, N. (2024) “Den KI-Reifegrad des Personalmanagements bestimmen”, *Controlling & Management Review*, 68(2), pp 50–55, https://doi.org/10.1007/s12176-023-1106-0.

Becker, J., Knackstedt, R. and Pöppelbuß, J. (2009) “Developing Maturity Models for IT Management”, *Business & Information Systems Engineering*, 1(3), pp 213–222, https://doi.org/10.1007/s12599-009-0044-5.

Boselie, P., van Harten, J. and Veld, M. (2021) “A human resource management review on public management and public administration research: stop right there…before we go any further…”, *Public Management Review*, 23(4), pp 483–500, https://doi.org/10.1080/14719037.2019.1695880.

Bruin, T. de, Rrosemann, M., Freeze, R. and Kaulkarni, U. (2005) “Understanding the Main Phases of Developing a Maturity Assessment Model”, In D. Bunker, B. Campbell and J. Underwood (Eds.), *Australasian Conference on Information Systems (ACIS)*, pp 8–19, https://eprints.qut.edu.au/25152/.

Chui, M., Hazan, E., Roberts. R., Singla, A., Smaje, K., Sukharevsky, A., Yee, L. and Zemmel, R. (2023) “The economic potential of generative AI. The next productivity frontier”, [online], https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

Deloitte. (2024) “AI Readiness & Management Framework (aiRMF)”, [online], https://www2.deloitte.com/content/dam/Deloitte/us/Documents/public-sector/ai-readiness-and-management-framework.pdf.

DFKI. (2022) “KI-Readiness-Check”, [online], https://werner.dfki.de/readinesscheck.

ElementAI. (2020) “The AI Maturity Framework. A strategic guide to operationalise and scale enterprise AI solutions”, [online], https://corysalveson.com/publication/eai-ramakrishnan-salveson-2020-ai-maturity-framework/EAI-Ramakrishnan-Salveson-2020-AI-maturity-framework.pdf.

Ellefsen, A.P.T., Oleśków-Szłapka, J., Pawłowski, G. and Toboła, A. (2019) “Striving for Excellence in AI Implementation: AI Maturity Model Framework and Preliminary Research Results”, *Logforum*, 15(3), pp 363–376, https://doi.org/10.17270/J.LOG.2019.354.

Eloundou, T., Manning, S., Mishkin, P. and Rock, D. (2023) “GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models”, *arXiv*, https://doi.org/10.48550/arXiv.2303.10130.

European Parliament. (2024) “EU AI ACT”, [online], https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

Franken, S. and Wattenberg, M. (2019) “The Impact of AI on Employment and Organisation in the Industrial Working Environment of the Future”, *Proceedings of the European Conference on the Impact of Artificial Intelligence and Robotics*, ECIAIR, pp 141–148, https://doi.org/10.34190/ECIAIR.19.096.

Fraser, P., Moultrie, J. and Gregory, M. (2002) “The use of maturity models/grids as a tool in assessing product development capability”, *IEEE International Engineering Management Conference,* pp 244–249, https://doi.org/10.1109/IEMC.2002.1038431.

Frick, N.R.J., Mirbabaie, M., Stieglitz, S. and Salomon, J. (2021) “Maneuvering through the stormy seas of digital transformation: the impact of empowering leadership on the AI readiness of enterprises”, *Journal of Decision Systems*, 30(2-3), pp 235–258, https://doi.org/10.1080/12460125.2020.1870065.

Fukas, P., Rebstadt, J., Remark, F. and Thomas, O. (2021) “Developing an Artificial Intelligence Maturity Model for Auditing”, *ECIS 2021 Research Papers (133),* https://aisel.aisnet.org/ecis2021_rp/133.

Gentsch, P. (2019) *AI in Marketing, Sales and Service*, Springer International Publishing. https://doi.org/10.1007/978-3-319-89957-2.

Greenhouse. (2023) “Concern about bias doesn't dampen HR professional' interest: Results from the EMEA HR Manager AI & Bias Pulse report”, [online], https://www.greenhouse.com/blog/results-from-the-emea-hr-manager-ai-bias-pulse-report.

Haenlein, M. and Kaplan, A. (2019) “A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence”, *California Management Review*, 61(4), pp 1–10, https://doi.org/10.1177/0008125619864925.

Hays. (2024) “HR-REPORT 2024. Schwerpunkt: Wie Künstliche Intelligenz die Unternehmenswelt beeinflusst”, [online], https://www.hays.de/documents/10192/118775/hays-hr-report-2024-kuenstliche-intelligenz.pdf.

Hazan, E., Madgavkar, A., Chui, M., Smit, S., Maor, D., Singh Dandona, G. and Huyghues-Despointes, R. (2024) “A new future of work: The race to deploy AI and raise skills in Europe and beyond”, [online],

https://www.mckinsey.com/mgi/our-research/a-new-future-of-work-the-race-to-deploy-ai-and-raise-skills-in-europe-and-beyond.

Holmström, J. (2022) "From AI to digital transformation: The AI readiness framework", *Business Horizons*, 65(3), pp 329–339, https://doi.org/10.1016/j.bushor.2021.03.006.

Jaaksi, J. (2018) "How to define an organisation's maturity to adopt artificial intelligence solutions [Master's Thesis]", *Turku School of Economics*, https://doi.org/10.13140/RG.2.2.15726.46401.

Jöhnk, J., Weißert, M. and Wyrtki, K. (2021) "Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors", *Business & Information Systems Engineering*, 63(1), pp 5–20, https://doi.org/10.1007/s12599-020-00676-7.

Kaushal, N., Kaurav, R.P.S., Sivathanu, B. and Kaushik, N. (2023) "Artificial intelligence and HRM: identifying future research Agenda using systematic literature review and bibliometric analysis", *Management Review Quarterly*, 73(2), pp 455–493, https://doi.org/10.1007/s11301-021-00249-2.

Kohlegger, M., Maier, R. and Thalmann S. (2009) "Understanding maturity models. Results of a structured content analysis", *Proceedings of I-KNOW '09 and I-SEMANTICS '09*, pp 51–61.

Kreutzer, R.T. and Sirrenberg, M. (2020) "AI Challenge—How Artificial Intelligence Can Be Anchored in a Company", In R.T. Kreutzer and M. Sirrenberg (Eds.), *Management for Professionals. Understanding Artificial Intelligence*, pp 235–273, Springer International Publishing. https://doi.org/10.1007/978-3-030-25271-7_10.

Kumar, M., Raut, R.D., Mangla, S.K., Ferraris, A. and Choubey, V.K. (2022) "The adoption of artificial intelligence powered workforce management for effective revenue growth of micro, small, and medium scale enterprises (MSMEs)", *Production Planning & Control*, pp 1–17, https://doi.org/10.1080/09537287.2022.2131620.

Lichtenthaler, U. (2020) "Five Maturity Levels of Managing AI: From Isolated Ignorance to Integrated Intelligence", *Journal of Innovation Management*, 8(1). https://doi.org/10.24840/2183-0606_008.001_0005.

Limat, C. (2022) "Disruptionspotenzial künstlicher Intelligenz: Ein Reifegradmodell zur Einführung ganzheitlicher KI-Initiativen in Unternehmen", *Wirtschaftsinformatik & Management*, 14(1), pp 60–67, https://doi.org/10.1365/s35764-021-00379-y.

Malik, N., Tripathi, S.N., Kar, A.K. and Gupta, S. (2022) "Impact of artificial intelligence on employees working in industry 4.0 led organisations", *International Journal of Manpower*, 43(2), pp 334–354, https://doi.org/10.1108/IJM-03-2021-0173.

McGraw, K.L. (1989) *Knowledge acquisition: principles and guidelines*, Prentice Hall.

MITRE. (2022). "The MITRE AI Maturity Model and Organizational Assessment Tool Guide: A Path to Successful AI Adoption", [online], https://www.mitre.org/sites/default/files/2023-11/PR-22-1879-MITRE-AI-Maturity-Model-and-Organizational-Assessment-Tool-Guide.pdf.

Palos-Sánchez, P.R., Baena-Luna, P., Badicu, A. and Infante-Moro, J. C. (2022) "Artificial Intelligence and Human Resources Management: A Bibliometric Analysis", *Applied Artificial Intelligence*, 36(1), Article 2145631. https://doi.org/10.1080/08839514.2022.2145631.

Pullen, W. (2007) "A public sector HPT maturity model", *Performance Improvement*, 46(4), pp 9–15, https://doi.org/10.1002/pfi.119.

Reichl, G. and Gruenbichler, R. (2023) "Maturity Models for the Use of Artificial Intelligence in Enterprises: A Literature Review", *19th International Scientific Conference on Industrial Systems,* pp 486–502, https://doi.org/10.24867/IS-2023-VP1.1-5_07341.

Rigamonti, E., Gastaldi, L. and Corso, M. (2024) "Measuring HR analytics maturity: supporting the development of a roadmap for data-driven human resources management", *Management Decision*, 62(13), pp 243–282, https://doi.org/10.1108/MD-11-2023-2087.

Saari, L., Kuusisto, O. and Pirttikangas, S. (2019) "AI Maturity Web Tool Helps Organisations Proceed with AI", *VTT White Pape*r, https://doi.org/10.32040/Whitepaper.2019.AIMaturity.

Sadiq, R.B., Safie, N., Abd Rahman, A.H. and Goudarzi, S. (2021) "Artificial intelligence maturity model: a systematic literature review", *PeerJ. Computer science*, 7, e661, https://doi.org/10.7717/peerj-cs.661.

Schuster, T., Waidelich, L. and Volz, R (2021) "Maturity Models for the Assessment of Artificial Intelligence in Small and Medium-Sized Enterprises", *Lecture Notes in Business Information Processing*, pp 22–36, https://doi.org/10.1007/978-3-030-85893-3_2.

Sonntag, M., Mehmann, S., Mehmann, J. and Teuteberg, F. (2024) "Development and Evaluation of a Maturity Model for AI Deployment Capability of Manufacturing Companies", *Information Systems Management*, pp 1–31, https://doi.org/10.1080/10580530.2024.2319041.

Stahl, B.C., Brooks, L., Hatzakis, T., Santiago, N. and Wright, D. (2023) "Exploring ethics and human rights in artificial intelligence – A Delphi study", *Technological Forecasting and Social Change*, 191, 122502, https://doi.org/10.1016/j.techfore.2023.122502.

Tambe, P., Cappelli, P. and Yakubovich, V. (2019) "Artificial Intelligence in Human Resources Management: Challenges and a Path Forward", *California Management Review*, 61(4), pp 15–42, https://doi.org/10.1177/0008125619867910.

Yams, N.B., Richardson, V., Shubina, G.E., Albrecht, S., and Gillblad, D. (2020) "Integrated AI and innovation management: The beginning of a beautiful friendship", *Technology Innovation Management Review*, 10(11), pp 1–18, https://doi.org/10.22215/timreview/1399.

Wendler, R. (2012) "The maturity of maturity model research: A systematic mapping study", *Information and Software Technology*, 54(12), pp 1317–1339, https://doi.org/10.1016/j.infsof.2012.07.007.

Balancing AI in SMEs: Overcoming Psychological Barriers and Preserving Critical Thinking

Madeleine Block
Institute for Applied Research for Skilled Crafts and Trades, Vienna, Austria

madeleineblock@gmx.net

Abstract: The rapid integration of artificial intelligence (AI) into various business sectors is transforming the operational landscape. This paper focuses on small and medium-sized enterprises (SMEs) and examines how they can achieve an optimal psychological balance when using AI. The goal is to encourage employees to adopt AI while maintaining critical thinking and autonomy. This research paper uses theoretical analysis as the research method, drawing on established psychological theories such as the Technology Acceptance Model and the Job Demands-Resources Model. The findings are used to propose a model for determining an optimal balance for SMEs. The analysis reveals several psychological barriers related to anxiety, which can lead to increased stress, lower motivation, and general resistance to the use of AI tools. Conversely, once the fear is overcome, there is a risk of over-reliance on AI. Therefore, it is important to provide training that helps employees recognize the benefits of AI and its impact on their tasks, critically evaluate AI recommendations, and find a balance between automated guidance with human judgement. Finding the optimal balance in the use of AI is critical. Fostering a culture of continuous learning, and adaptability, together with supportive leadership, can help maintain this balance. In summary, through a strategic application of psychological theories, SMEs can harness the potential of AI to improve work performance and mitigate labour shortages while developing motivated and critically thinking employees.

Keywords: Artificial intelligence, Technology acceptance, Psychological barriers, Critical thinking, Organisation, Model

1. Introduction

This paper examines how implementing artificial intelligence (AI) can succeed in organisations. Psychology as the science of human experiences and behaviour is embedded in the context of the digital working world. Specifically, the interaction between people and organisations in dealing with AI is examined. The key question is: How can organisations design a successful implementation so that AI is perceived as a relief rather than a burden and can contribute to an increase in productivity for the entire organisation?

Since the so-called "first industrial revolution" in the second half of the 18th century, people have been accompanied by technical achievements that generally cause discomfort at first. In Goethe's "Wilhelm Meisters Wanderjahre", textile machines threaten human activity, and Nietzsche speaks of the "idleness" triggered by the machine. Even today, in the so-called "digital age" or "age of artificial intelligence" (Castells 2010, Russel and Norvig 2022) people and organisations are grappling with the question of how to approach the new or avoid it.

Organisation is a multi-layered term, and the understanding of the term varies depending on the discipline. In organisational psychology, organisations are generally understood as social systems (Dorsch 2024, i.e. a complex network of technical and social structures. Technological developments influence the organisation of work and the social structure of employees and vice versa.

In the second chapter, findings and background information on technological progress and AI are presented to create an understanding of the organisational psychological spectrum of AI. Three relevant theories are then considered to present an integrative model based on their findings in the fourth chapter, which can contribute to the acceptance and successful introduction of AI by taking organisational, technological and psychological factors into account. Then, an interlinkage to preserving critical thinking is made before the main points of this work are summarised in the fifth chapter.

2. Background

Technological progress goes hand in hand with a reduction in the workload for people. The structural change from an agricultural to an industrial and service society to a digital society is one of the reasons for a high level of social mobility, both in the vertical direction of career advancement and relegation and in the horizontal direction of changing places of work and residence (Castells 2010; Ebeling et al. 2012). AI also provides support where it can relieve the burden on people (Langer et al. 2021). Since the introduction of machines, working hours have fallen from 72 hours to less than 40 hours per week without any major impoverishment, but on the contrary an increase in living standards and life expectancy (Statista 2024, 2024b). Technical innovations should therefore not cause any unease. But when something is difficult to understand, it tends to make people feel uneasy or indifferent (Tsao et al. 2013). Increasingly complex organisational structures, transnational production and distribution chains as well as AI software serve precisely this impenetrability. Technical progress with ever-

shorter innovation cycles means that the goal of "being finished" can hardly be achieved, which brings with it a considerable degree of uncertainty. However, it is precisely the knowledge of recurring change that may facilitate the process and is made possible by human flexibility and adaptability.

The rapid technical progress of AI with its presumed productivity-enhancing potential, as well as the increasing shortage of skilled labour, have rekindled interest in AI in science and practice in recent years. For instance, the latest survey results in Germany from 2023 show that around 12% of German companies are already using at least one AI software, around 10% are planning to implement one and just over 35% are holding talks about potential AI use cases (OECD 2024). The main obstacles to the hesitant implementation of AI in the organisational context, especially among SMEs, include insufficient digitalisation - which is a prerequisite for the efficient use of AI, an insufficient amount of qualitatively prepared data, a poor understanding of the potential benefits of AI, uncertainty regarding data protection aspects and a lack of AI skills (OECD 2024).

A transformation towards AI-using organisations requires a rethink and a holistic approach, as AI has an impact on a variety of organisational areas. At the same time, care should be taken to ensure that employees perceive AI as increasingly attractive and remain loyal to the organisations.

Artificial Intelligence

"Intelligence" plays an important role in psychology and intelligence tests are one of the most successful diagnostic procedures. Although there are many views on what is meant by intelligence, which is reflected in the large number of intelligence models and correspondingly different intelligence tests, a relatively standardised understanding of what intelligence is has emerged. Intelligence represents both a personal characteristic and a procedural meaning, as it describes how something is done: quick comprehension, in-depth understanding, precise judgement, learning from experience, etc. (Schmidt-Atzert et al. 2021, Myers 2014).

AI goes beyond robotics and human-machine interaction and includes data science, affective computing, virtual realities (Hasenbein 2023, Langer et al. 2021). The definition of AI has changed over time. AI research began in the 1950s intending to simulate human-like thinking. Alan Turing introduced the Turing Test in the 1950s with the key question: Can a computer converse like a human? He assumed that a computer has intelligence on par with a human if a third party cannot distinguish who the human dialogue partner is (Turing 1950). As early as the 1960s, Joseph Weizenbaum's ELIZA program was one of the first computer programmes to conduct simple conversations with users, briefly giving the impression that a human was communicating (Weizenbaum 1966). Another founding father of AI is John McCarthy, who coined the English term "Artificial Intelligence" in civilian usage and surmised that any aspect of learning or any other characteristic of intelligence can in principle be described so precisely that a machine can simulate it (McCarthy et al. 2006). Today, the definition of AI focuses on the ability of machines to perform cognitive tasks efficiently and autonomously, reflecting advances in machine learning and neural networks (deep learning). An advanced subgroup is generative AI, which focuses on the creation of new content such as images, text, videos, music, etc. It aims to recognise and imitate patterns in existing data to generate new results (Kulkarni et al. 2023).

The rise of AI since the 2010s is mainly due to the availability of data, the further development of algorithms, the improvement of hardware and software and the increase in investment capital (Holdren et al. 2026, OECD 2024). Since the freely accessible publication of ChatGPT on 30 November 2022, AI has become a much more popular and becomes increasingly integrated into various sectors of daily working life, from decision-making to automating tasks.

The changing definition of AI reflects the growing capabilities and expectations of the technology. From an organisational psychology perspective, it is important to understand this development to promote the acceptance of AI in organisations. Employees must not only be trained in the use of AI but also develop an understanding of its potential. Successful integration of AI therefore requires both technological adjustments and a change in values towards greater openness and willingness to learn.

3. Theoretical Concepts

Research questions in the field of AI have traditionally been very technical and are increasingly becoming application oriented. AI applications, i.e. AI software, are a young business field and the acceptance of AI by users is still little researched. New theories and practical tools are needed for dealing with AI and its implementation in organisations. This article aims to develop a preliminary model that enables a holistic view, including technological, psychological and emotional factors, embedded in the organizational context. This is done based on three established scientific concepts, which are presented below.

3.1 7-S Model

The 7-S Model, developed in the 1980s by McKinsey consultants Waterman, Peters and Phillips, is a standard framework in business and management literature for analysing organisations that incorporates seven variables that interact with each other. The variables are divided into hard factors: 1. structure, 2. strategy, 3. systems and soft factors: 4. skills, 5. staff, 6. style and 7. superordinate goals (Waterman et al. 1980). A particular contribution of this model is the emphasis that structure alone is not enough, but that a holistic approach is necessary and an awareness that organisations adapt slowly to change (Waterman et al. 1980). This integrated approach enables organisations to understand and manage the complexity of organisational change, especially when it comes to implementing new technologies such as AI. The 7-S Model offers a top-down, strategic organisational perspective and framework that ensures that all aspects of the organisation are considered in the acceptance process.

3.2 Technology Acceptance Model

Fred D. Davis developed a technology acceptance model (TAM) with two variables that are fundamental determinants of user acceptance of information technology: 1) perceived usefulness and 2) perceived ease of use. The regression analyses indicate that perceived ease of use may be a causal precursor of usefulness. These findings are significant as they provide validated measures for predicting user acceptance of technology (Davis 1989). Davis and Venkatesh extended the TAM to include additional factors that influence perceived usefulness. Firstly, subjective norms and image from theories of social influence, and secondly, job relevance, output quality and result demonstrability from theories of cognitive instrumental processes (Venkatesh and Davis 2000). The resulting TAM2 also takes user experience and voluntariness into account, which allows the acceptance of new technologies in an organisational context to be better captured. In a further study, Venkatesh and Bala combine the TAM2 model and the determinants of perceived ease of use to form TAM3 (Venkatesh 2000). It is hypothesised that individuals develop an initial assessment of perceived ease of use through various anchors of their attitudes towards computers and their use. These anchors include computer self-efficiency, i.e. the belief in one's ability, the perception of external control, i.e. the belief in the availability of necessary structures and resources, computer anxiety and computer playfulness as an intrinsic motivation to try new things (Venkatesh and Bala 2008). When practical experience has been gained through actual use, perceived enjoyment and objective usability influence perceived ease of use.

The model provides the individual and experiential perspective as well as consistently explains about 40% of the variance in individuals' intention to use an IT and actual usage (Venkatesh and Bala 2008). The added value of TAM3 is that psychological and emotional factors are considered, allowing the model to better capture the diversity of human reactions to new technologies.

3.3 Job Demands-Resources Model

The Job-Demands-Resources model (JD-R) was introduced in 2001 and has since developed into a comprehensive theory for analysing and improving working conditions and employee well-being (Demerouti et al. 2001, Bakker and Demerouti 2017). The JD-R model provides organisations with a structured framework to understand the complex interactions between job demands and job resources and to analyse their impact on employee wellbeing and performance. Alongside stress models such as the Demand-Control Model and the Effort-Reward Imbalance Model, the JD-R Model is more flexible as it considers a broader range of working conditions and incorporates both negative and positive indicators of employee well-being (Bakker and Demerouti 2007). This flexibility is particularly valuable for the introduction of AI, as it brings completely new requirements and resources into the working environment.

The model divides all working conditions into two main categories: Job demands and job resources. While job demands encompass aspects of work that require sustained physical or mental effort and are therefore associated with certain physiological and psychological costs, job resources serve to achieve work goals, reduce work demands or promote personal growth and development (Bakker and Demerouti 2017). The JD-R model is based on two different processes. One is the health-impairing process, which can lead to exhaustion and health problems if work demands are high. The second is the motivational process, which can lead to increased commitment and better performance based on job resources. An important aspect is that job resources can buffer the negative impact of high job demands on stress (Bakker and Demerouti 2017). The JD-R model enables organisations to develop targeted interventions to minimise the negative effects of high demands and maximise the positive effects by providing appropriate resources.

This model bridges the individual and organisational levels by considering as it considers the balance between job demands and resources, which is crucial when introducing new technologies like AI. It helps in understanding how the introduction of AI might affect employee well-being and performance.

4. Integrative Model for AI Implementation

The integrative model for AI implementation shown in Figure 1 is a first attempt to analyse and implement the introduction of AI in organisations. It was developed based on elements of the TAM3, the JD-R model and the 7-S model and is discussed below. The combination of these three frameworks allows for a holistic approach to AI adoption and helps to overcome the limitations of using any single model. For instance, while TAM is excellent at predicting individual acceptance, it doesn't consider organisational factors. The 7-S Model fills this gap. Similarly, the JD-R model helps to understand the impact of AI adoption on employee well-being and performance, which neither TAM3 nor the 7-S Model directly addresses.

This integrative model considers technological as well as psychological and emotional aspects of AI adoption. It addresses individual factors such as fears and self-confidence, which are often barriers and embeds them in a broader organisational context. The model emphasises the need to align all factors, which is crucial for a successful implementation of AI.

Organisations are embedded in an *environment* with which they interact, e.g. AI developments affect them, but also organisations affect AI. Within the organisation, the organisational factors (cf. 7-S model) form the framework for the introduction of AI; they interact with each other as well as with individuals and groups, which in turn influence individual and organisational behaviour. An organisation's strategy should include the increased use of AI. The structure and the system should be flexibly adapted to the AI software, an appropriate IT infrastructure should be created, and employees should be trained to develop the necessary skills in dealing with AI.

Care should be taken to ensure that the style of management is characterised by greater flexibility to support new working methods. The role of AI should also be reflected in values such as willingness to learn, tolerance of mistakes and openness to new developments. As already made clear in the 7-S model, an organisation can only achieve its economic goals if it takes people seriously as human beings. While introducing AI, this is likely to be a particularly important point.

Good external conditions (external control), i.e. resources and support options made available to a person by the organisation, increase the likelihood that AI will be perceived as easy to handle. Provided resources can also support individual self-confidence and strengthen trust in one's abilities in dealing with AI software, e.g. through positive experiences. The strategy and working conditions can also inhibit or increase the fear of AI. An increased level of anxiety often leads to a perceived burden/stress and can result in the development of an aversion to AI. On the other hand, the playful element of AI can be emphasised, which should be possible through user-friendly tools on the market with appropriate guidance. By taking these soft factors into account, the model does better justice to the diversity of human reactions to new technologies. Targeted measures for understanding and promoting the acceptance of AI can overcome both technological and psychological barriers.

At the centre of this model is the perceived benefit, i.e. managers and employees must understand how AI can improve their work performance and productivity. In addition, the AI software used must be perceived as user-friendly and accessible to people with different levels of technical knowledge. Building on this, workshops and communication strategies can be developed that emphasise the usability of AI and demonstrate its user-friendliness to reduce resistance and increase the acceptance of AI.

AI software brings new tasks for the organisation and the employees at whose workplaces it is used, and in any case, those affected must adapt. This means that they must forget some of their old ways of working to learn new things. During AI implementation, new work requirements such as training, adaptation of the system to the AI software, etc. and new work resources through AI such as the automation of routine tasks, the generation of texts, images, ideas, etc. must be identified. This helps to understand how AI can be both stressful due to its increased opacity and motivating due to new possibilities and increased utility. Knowledge of the buffer effect of resources should be used to provide targeted resources such as clear communication, application-related support, etc. to minimise potential stress during the implementation of the AI strategy.

Particularly in the initial phase of implementing AI software, care should be taken to ensure that individuals are not overwhelmed by new tasks or an increased workload but are provided with resources such as practical assistance with relevant AI software that they can use effectively for their work. The challenge of our time lies

in the fast pace of life, in the sensory overload, so care should be taken not to overburden individuals and to adapt the pace of the flood of information to their capacities. Collaboration and the exchange of experience among employees and in working groups should also be encouraged when adapting to AI technology.

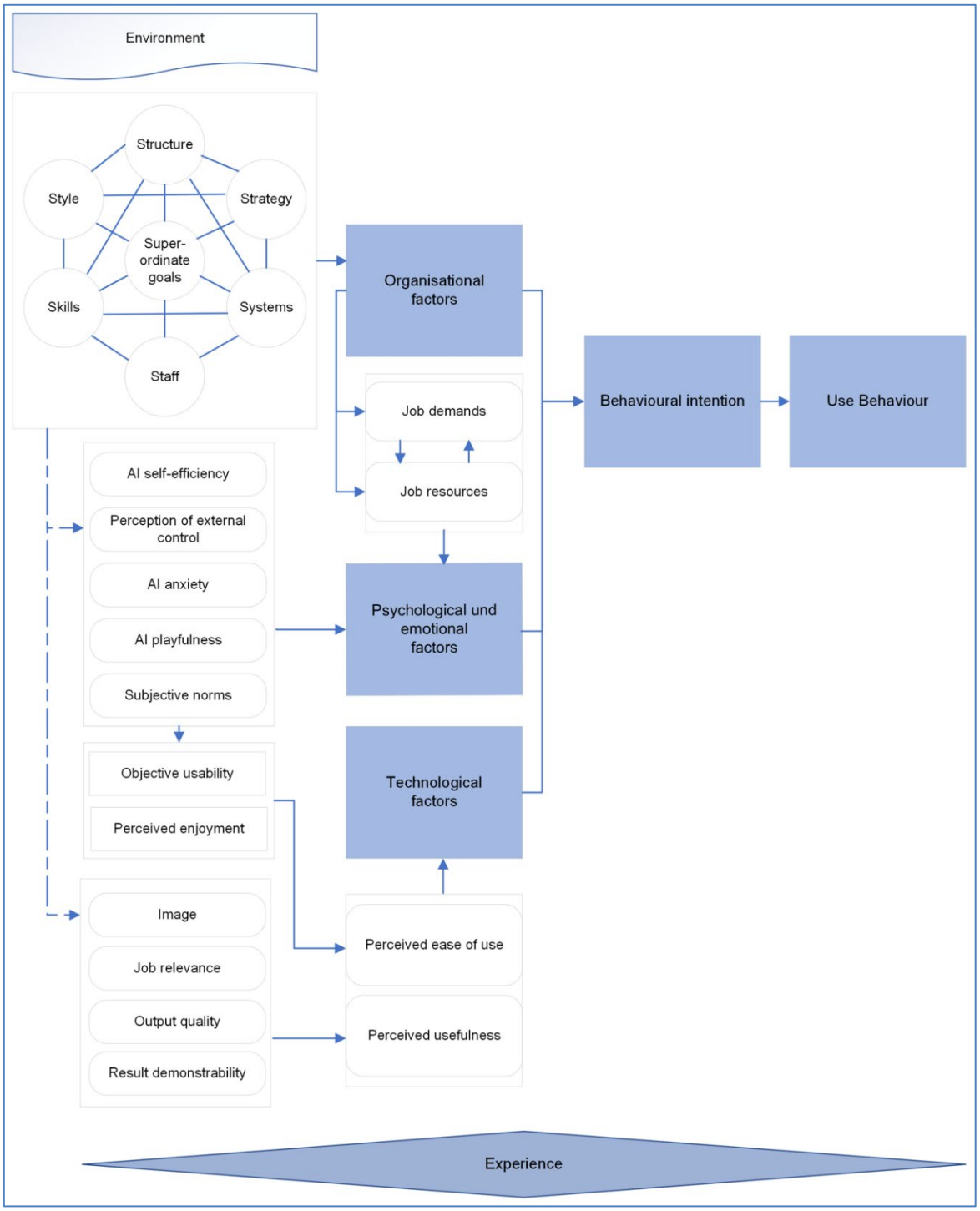


Figure 1: Integrative model for AI implementation

There are relationships between the factors and events in the model. These relationships cannot generally be fully recognised and predicted by humans; they are "probability relationships". During the changeover, work performance will generally drop initially and only gradually reach a new level of productivity. The critical moment is the changeover/adaptation to new environmental conditions, and management is well advised to make employees aware of initial difficulties, but also of the fact that successes can be expected subsequently, and to initiate the motivational process and overcome psychological barriers through positive experiences. As with all activities and technological innovations, human control is required. The results of AI should be

critically scrutinised by the organisation and everyone. On the one hand, this serves to prevent errors and, on the other, it gives people meaning.

5. Preserving Critical Thinking

Assuming employees have accepted AI successfully, the user-friendliness and usefulness of the AI tools have been recognised, as they experience the benefits of AI in their daily work. This raises the question, can't it be that employees and humans are easily inclined to trust in the work of AI and no longer question it, to sit back and stop thinking?

Taken this idea further, means there is a growing concern about over-reliance on AI, leading to the "lay-back effect," where business leaders and employees become passive and uncritically accept AI-generated outputs. This dependence can in turn impede innovation, diminish problem-solving skills, and expose organisations to errors and biases inherent in AI systems. Exactly what should be increased. That's why maintaining critical thinking is inherently important.

Critical thinking begins with questioning assumptions and examining the evidence behind claims. Browne and Keeley (2015) argue that this skill is especially important in today's information-saturated world, where individuals are often bombarded with data, opinions, and conflicting viewpoints. A critical thinker actively engages with information by identifying biases, recognizing logical fallacies, and seeking out reliable sources. For instance, when evaluating news reports or online information, a critical thinker would not only consider the source but also question its motivations and underlying assumptions.

A significant risk associated with over-reliance on AI is the potential suppression of creativity and innovation. SMEs often thrive on their capacity for creative thinking, niche market identification, and rapid adaptation to change. However, when decision-making becomes overly dependent on AI tools, employee disengagement from the creative process may ensue, leading to a stagnation of innovative ideation. AI systems are designed to analyse existing data patterns but are not inherently equipped to engage in "out-of-the-box" thinking or generate entirely novel concepts (Tegmark, 2017). This creates a potential trap where businesses, instead of thriving on fresh ideas, settle into routines dictated by AI-driven patterns.

Although AI offers significant opportunities to enhance operational efficiency and competitiveness, it is crucial to avoid becoming overly dependent on these technologies. The "lay-back effect," where employees blindly trust AI recommendations, can erode critical thinking, limit creativity, and expose businesses to risks from biased or flawed AI outputs. To remain agile and competitive, SMEs must maintain a balance between leveraging AI for efficiency and fostering a culture that prioritizes human insight, critical thinking, and creative problem-solving.

How can employees in SMEs create this balance? It is essential to develop an awareness of the questioning of results, i.e. regularly review evaluations of AI-generated results. For example, a marketing expert can critically question the recommendations of an AI-driven targeting tool by incorporating their own market research data and customer feedback. This helps to identify potential errors or biases in the AI predictions and to develop alternative approaches. Another way is working together in interdisciplinary teams that include different departments such as marketing, sales and product development. This can also bring different perspectives to the table and encourages creative thinking, helping to develop innovative solutions that go beyond AI recommendations. Regular brainstorming sessions can be used to share ideas, test applications, and implement feedback mechanisms. A weekly or monthly exchange about the AI results and their impact on work can be helpful. These discussions allow employees to be more aware of the results and critically consider how to optimise the use of AI. One example would be internal meetings, in which employees show each other their experiences with AI in very specific use cases, sharing benefits and concerns. Such feedback can also help management improve AI tools and ensure that the human perspective is integrated into the decision-making process. Employees can learn to think of AI as another supportive employee and to treat him as such with all the advantages and disadvantages.

In doing so, SMEs ensure that AI remains a tool that complements human judgment, rather than replacing it.

6. Conclusion

In this paper, AI was examined from an organisational psychology perspective to better understand what AI is and how AI can be successfully implemented in organisations.

The starting point for the considerations was that organisations are social systems capable of learning. Section 2 looked at the period from industrialisation to the digital age in terms of technological developments. It was

shown that the goal of making machines a reflection of humans and, in this sense, AI is not just a contemporary issue. Nor does today's knowledge point to a technical reconstruction of the human being. Nevertheless, generative AI has reached a milestone that is changing the world of work and organisations significantly. This presents organisations with major challenges in dealing with AI and the need to retrain. AI is giving job profiles new job characteristics and may also lead to job losses. Every cultural innovation, including AI, harbours opportunities in a positive sense, but also risks. It can be helpful to focus on the positive opportunities and familiarise oneself with the new so that it becomes more familiar, and organisations can learn to deal with or avert the risks. Section 3 introduced theoretical concepts that are useful to the research question and, building on this, Section 4 presents a first attempt at an AI implementation model. The model offers a holistic approach that takes organisational, technological and psychological factors into account and stimulates an iterative learning process based on gaining experience and experimenting with AI. It's an attempt to bridge the gap between different levels of analysis and provide a more complete picture of the AI implementation process in organisations. In the 5th section, an outlook was ventured and the question was raised of what happens when AI Tools are accepted and used by employees, the added value is experienced. In this context, the importance of maintaining critical thinking also concerning AI was emphasized.

The present work can only be a starting point for the exploration of organisational diagnoses and concrete design approaches for the implementation of AI in organisations (Hasenbein 2023). In further empirical research, the analysis of the individual factors of the model could be used to identify psychological and organisational barriers and to plan measures for targeted interventions to overcome identified barriers and to accompany the concrete implementation of AI, taking all model elements into account. During the process, the behavioural factors and production increases would be regularly evaluated and the strategy adjusted if necessary. Another interesting approach could be the inclusion of findings from stress theory, e.g. that the degree of aversiveness of a stimulus is not only defined by intensity, novelty and duration but also by the coping options and their prospects of success and failure (Lazarus and Folkman 2015). If organisations are aware of this, they can strengthen their own adaptability and that of individual employees by examining the environment to determine the extent to which aversive events are predictable and what options exist for influencing them or actively or passively avoiding them.

Research in the field of AI shows that the contribution of psychology could have a decisive influence in the future, both in counselling and in basic research, for example in the field of affective computing.

References

Bakker, A. B. and Demerouti, E. (2007) “The Job Demands-Resources model: state of the art”, *Journal of Managerial Psychology*, Vol. 22, No. 3, pp. 309–328.

Bakker, A. B. and Demerouti, E. (2017) “Job demands-resources theory: Taking stock and looking forward”, *Journal of Occupational Health Psychology*, Vol. 22, No. 3, pp. 273–285.

Browne, M. N. and Keeley, S. M. (2015) *Asking the right questions: A guide to critical thinking*, 11th ed. Pearson.

Castells, M. (2010) *The Rise of the Network Society*, 2nd ed., Blackwell, Oxford.

Davis, F. D. (1989) “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”, *MIS Quarterly*, Vol. 13, No. 3, pp. 319-340.

Demerouti, E., Bakker, A. B., Nachreiner, F. and Schaufeli, W. B. (2001) “The job demands-resources model of burnout”, *Journal of Applied Psychology*, Vol. 86, No. 3, pp. 499–512.

Dorsch, Lexikon der Psychologie (2024) *Organisation*, Markus Antonius Wirtz, [online], Hogrefe AG, https://dorsch.hogrefe.com/stichwort/organisation.

Ebeling, I., Vogelauer, W. and Kemm, R. (2012) *Die Systemisch-dynamische Organisation im Wandel. Vom fließenden Umgang mit Hierarchie und Netzwerk im Veränderungsprozess*, Bern, Stuttgart, Wien, Haupt Verlag.

Hasenbein, M. (2023) *Mensch und KI in Organisationen*, Berlin, Heidelberg, Springer.

Holdren, J. P., Bruce, A., Felten, E., Lyons, T. and Garris, M. (2016) *Preparing for the future of artificial intelligence*, Executive Office of the President National Science and Technology Council Committee on Technology. Washington, D.C.

Kulkarni, A., Shivananda, A., Kulkarni, A. and Gudivada, D. (2023) *Applied Generative AI for Beginners*, Berkeley, CA, Apress.

Langer, M., Bajwa, N. H. and König, C. J. (2021) *Arbeits- und Organisationspsychologie im 21. Jahrhundert*, Wiesbaden, Springer Fachmedien Wiesbaden.

Lazarus, R. S. and Folkman, S. (2015) *Stress, appraisal, and coping*, 11th ed., New York, Springer.

McCarthy, J., Minsky, M. L., Rochester, N. and Shannon, C. E. (2006) “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence”, *AI Magazine*, Vol. 27, No. 4.

Myers, D. G. (2014) *Psychologie*, 3rd ed., Berlin, Springer.

OECD (2024) *OECD-Bericht zu Künstlicher Intelligenz in Deutschland*, Paris, OECD Publishing.

Russell, S. J. and Norvig, P. (2022) *Artificial intelligence. A modern approach*. 4th ed., [online], Pearson, https://elibrary.pearson.de/book/99.150005/9781292401171.

Schmidt-Atzert, L., Krumm, S. and Amelang, M. (2021) *Psychologische Diagnostik*, 6th ed., Berlin, Heidelberg, Springer.

Statista (2024a) *Entwicklung der Lebenserwartung von einjährigen Mädchen und Jungen in Deutschland in den Jahren von 1871 bis 2017*, [online], Statista Research Department, https://de.statista.com/statistik/daten/studie/1127745/umfrage/ lebenserwartung-von-neugeborenen/.

Statista (2024b) *Wochenarbeitszeit in Deutschland in den Jahren 1871 bis 1990*, [online], Statista Research Department, https://de.statista.com/statistik/daten/studie/1126144/umfrage/woechentliche-arbeitszeit-in-deutschland/.

Tegmark, M. (2017) *Life 3.0: Being human in the age of artificial intelligence*, Knopf Doubleday Publishing Group.

Tsao, A., Moser, M. and Moser, E. (2013) "Traces of experience in the lateral entorhinal cortex", *Current biology*, Vol. 23, No. 5, pp. 399–405.

Turing, A. M. (1950) "Computing Machinery and Intelligence", *Mind LIX*, Vol. 236, pp. 433–460.

Venkatesh, V. (2000) "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model", *Information Systems Research*, Vol. 11, No. 4, pp. 342–365.

Venkatesh, V. and Bala, H. (2008) "Technology Acceptance Model 3 and a Research Agenda on Interventions", *Decision Sciences*, Vol. 39, No. 2.

Venkatesh, V. and Davis, F. D. (2000) "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies", *Management Science*, Vol. 46, No. 2, pp. 186–204.

Waterman, R. H., Peters, T. J. and Phillips, J. R. (1980) "Structure is not organization", *Business Horizons*, Vol. 23, No. 3, pp. 14-26.

Weizenbaum, J. (1966) "ELIZA - a computer program for the study of natural language communication between man and machine", *Commun. ACM*, Vol. 9, No. 1, pp. 36–45.

The new era of Technology Mysticism: Generative Artificial Intelligence and its Effects

Karsten Böhm[1] and Jürgen Sammet[2]
[1]FH Kufstein Tirol – University of Applied Sciences, Kufstein, Austria
[2]HAM – University of Applied Management, Germany

Karsten.Boehm@fh-kufstein.ac.at
JuergenSammet@fhham.de

Abstract: Generative Artificial Intelligence (GenAI) started to disrupt many application areas in the domain of information technology and is developing at a rapid pace. GenAI exhibits different systemic characteristics, it is a trained technology as opposed to the engineered technologies that have been developed in the IT domain in the past. As trained technologies are heavily depending on huge amounts of training data, their behaviour is not deterministic but of a stochastic nature, leading to a limited understanding of those systems. their behaviour is emergent. Assumptions and beliefs in the abilities of technologies by their users create a new age of mysticism that can be compared with our past and the relation of people in medieval times to the age of Mysticism with respect to their understanding of nature and their surroundings. We are facing AI today like our ancestors faced incomprehensible natural phenomena. This article is discussing the resulting effects from a technical but also from a philosophical perspective.

Keywords: Artificial intelligence, Technology adaptation, Mysticism, Large Language Models, Generative technologies

1. Introduction

Our modern world is already crammed with technology, and our daily routines and many parts of the economic ecosystem are dependent on networked information technology. We are already used to live in that symbiosis to a certain extent and only realize our dependency when technology fails on the individual level, on the infrastructure level or on a service/product level.

Despite being complex technologies and technical systems, up to now those complexities and functionalities are engineered technologies, meaning that they are designed and built by humans with a detailed understanding about the inner workings of each and every component, following the general principle of building large and complex machinery out of smaller components that are simple enough to understand, to design and to build. Although we are aware that complex systems sometimes exhibit behaviour that is hard to explain the general assumption is that engineered technologies *are understandable, the behaviour is deterministic, controllable and reliable*. Because of those properties, we trust those systems and rely on them. This dependency got stronger due to the accelerating process of a digital transformation in the economy and the society.

With the advent of Generative Artificial Intelligence (GenAI), these assumptions are substantially challenged, mostly unnoticed. While the GenAI systems are still engineered systems with regard to their basic building blocks, *their behaviour is emergent*, driven by massive amounts of training data that is used to derive a numeric model with billions of parameters. And while this process is deterministic and algorithmic, the resulting behaviour is of a stochastic nature. These *trained technologies* are fundamentally different from the technologies mentioned above, as they cannot be understood in the same way. Despite their limited explainability and technology weaknesses, users and organizations are applying them increasingly for a wide area of use-cases, trusting in technologies that they do not (fully) understand.

Assumptions and beliefs in the abilities of technologies by their users create a new age of mysticism that can be compared with our past and the relation of people in medieval times to the age of Mysticism with respect to their understanding of nature and their surroundings. We are facing AI today like our ancestors faced incomprehensible natural phenomena: We currently do not really understand what AI is, what effects AI currently has or will have in the future.

This has a wide variety of effects: Myth-making, fear, subjugation, general rejection, etc. Those who actively use AI sometimes misjudge its possibilities and limitations. This paper tries to shed light on those phenomena and relate them to the lessons learned from the past – both from a technical but also from a philosophical perspective.

As in the past (the Age of Enlightenment) education might be a way to address this issue. Building competences in different domains lead to informed decisions and usage: technical (How does AI work?), cognitive (How can AI be evaluated?) and also emotional (How to deal with uncertainties etc.?). As we are in the process of this

relatively new technology domain, there is a high development dynamic that will lead to new insights and learnings on every level: individual, economic and societal.

2. Review of Mysticism in History and Technology Development

2.1 Myth, Enlightenment and Technique

Technologies are accompanied by a mix of truths and misunderstandings about their capabilities and limitations. They evoke both fascination and fears, resulting in both glorification and condemnation. Therefore, technologies are often mystified. This is not a modern phenomenon: the interplay between myths and technology can be traced far back in human history. A well-known example of this connection is the legend of Talos. Talos was a giant bronze automaton created by Hephaestus, the god of blacksmithing. His task was to protect the island of Crete. To do this, he circled the coastline three times daily, defending it by hurling massive rocks at intruding ships or by heating himself and killing enemies with an embrace. This early mythological concept of robots and artificial life illustrates the deep-rooted connection between myths and technological visions (Mayor 2018).

Myths are narratives that help humans explain natural phenomena or cultural practices. For the German philosopher Hans Blumenberg, myths are an anthropological constant: they aid in making an environment perceived as uncertain, mysterious and threatening more understandable (Blumenberg 1985). The story of Talos is a fitting example. The idea of a giant hurling rocks could have been an early interpretation of volcanic activity. This demonstrates how myths often served as explanatory models for natural events before scientific explanations were available. Such stories reflect human fears and desires and provide explanations for inexplicable phenomena. Myths function as narrative tools that allow people to address and rationalize their deepest existential questions. They are a means of gaining control over the unpredictability of nature and life. Through repeated telling and cultural embedding, myths gain significance and influence collective memory. Myths originated to provide answers to natural events that were not understandable. From a societal perspective, myths also served to legitimize existing power structures.

With the emergence of ancient philosophy, myth began to create a tension with the logos, the operative principle of the world. For Plato, philosophy had to overcome traditional myth through logos, questioning mythic narratives and replacing them with rational argumentation. Aristotle made a clear distinction between mythos and logos. For him, mythos had a poetic and narrative function, while logos held scientific and rational significance. This differentiation helped to solidify the role of logos in philosophy as a tool for gaining knowledge and explaining the natural world.

In the European Enlightenment, the tension between logos and mythos reached its culmination. The Enlightenment opposed a system of thought and society permeated by myths and uncritically followed authorities. In his famous essay "Answering the Question: What is Enlightenment?" Immanuel Kant defined Enlightenment as the "emergence of man from his self-imposed immaturity" (Kant 1991) and emphasized the importance of autonomy and the freedom to use reason publicly. This immaturity is self-imposed because people prefer to rely on authorities like priests and kings rather than thinking for themselves. It is not that people lack the intellectual capacity to think independently, but that they do not dare to do so. Thus, for Kant, the motto of the Enlightenment is "Sapere aude - Have the courage to use your own understanding" (Kant 1991). Enlightenment is demystifying. Compared to the mystically influenced medieval worldview, the Enlightenment represented a gain in reality, likely even a reality shock, due to the inevitable challenges that come with an unvarnished, unsparing view of the real world.

In the Age of Enlightenment, thinkers hoped that the application of reason would bring about the end of myths. However, this hope was ultimately disappointed, as the relationship between myth and logos is not a simple linear one but a complex and dialectical process. In "Dialectic of Enlightenment," Horkheimer and Adorno show that Enlightenment itself turns into a new form of myth. Rationality transforms into an all-pervasive instrumental rationality and gives rise to the myth of control and mastery. Similarly, proponents of postmodernism such as Richard Rorty and depth psychology such as Sigmund Freud and Carl Gustav Jung challenge the notion that modernity has overcome myth, emphasizing the "myth of mythlessness".

Myths continue to exist despite rationality and enlightenment. As mentioned, for Blumenberg, myth is an anthropological constant. Modern psychology has shown that we are not only rational beings but also non-rational and influenced by the unconscious. Myths seem to have their origins in this part of our psyche. Therefore, it is useful to distinguish between two levels of "myth."

The first level is characterized by misunderstandings and uncertainties that are initially not understandable but can be resolved through rationality. For example, thunder is no longer seen as a manifestation of divine wrath but as a natural event explainable by causal principles. This shift from mythical to scientific explanations marks the success of the natural sciences and promotes critical thinking. Rational Enlightenment explains what is empirically the case.

The second level contains mythical narratives—stories that cannot be fully replaced by rational explanations. Myths are more than just irrational views; they are complex narratives that help individuals and societies comprehend the world around them. These stories are imbued with symbolic meaning and cannot easily be displaced by scientific knowledge. Instead, they must be recognized and understood as myths. Often, these myths are shaped by deep-seated fears and hopes and serve as a cultural framework to grasp the mysteries of existence. One of the most powerful myths revolves around themes of salvation and damnation, heaven and hell.

2.2 Myth and Technique

This process of enlightenment and myth also accompanies the relationship between technology and humans. Traditionally, technology is considered a means to an end. However, technology becomes a myth when it is not only a means to an end but also part of a narrative of salvation. As "deficient beings", humans have always viewed technological developments with exaggerated expectations and visions of salvation. As seen in the example of the bronze giant Talos, such notions are found even in antiquity. The stories of Pandora and Pygmalion also illustrate such early technological myths of creating artificial life by humans.

These ancient myths often serve as cultural and intellectual forerunners for modern technological developments. They demonstrate that the human imagination and scientific impulses to use technology to create artificial life have existed for thousands of years. These narratives also contained ethical concerns and warnings about the potential dangers and abuses of such technologies. For example, the myth of Pandora indicates that technological gifts can be both blessings and curses, highlighting the ambivalent nature of technology (Mayor 2018).

This ambivalence of fascination and fear is also evident in many other technologies. Often, the myth of salvation prevails: for example, the invention of the steam engine in the 19th century or nanotechnology at the end of the last century were celebrated as solutions to all of humanity's major problems. In such narratives, technology is endowed with a quasi-religious salvific character, and engineers have been stylized as priests of the technological age since the end of the 19th century, fulfilling the longing for deliverance from the ambivalences of life.

In addition to the fundamental narrative of salvation and damnation, two other myths characterize technologies:

- The Myth of Progress: Since the 19th century, there has been a belief that technological innovations inevitably lead to societal progress. This belief was reinforced by the intro-duction of technologies such as the telephone, which was considered revolutionary even though it ultimately reflected existing social norms and behaviours.
- The Myth of Technological Determinism: The notion that technology is an autonomous force, acting independently of human influence and inevitably changing society, has solidified over time.

These myths contribute to the complex relationship between society and technology, where advancements are often viewed through a lens of both hope and apprehension.

2.3 Myth and Artificial Intelligence

Given this context, it is not surprising that the discussion around AI is heavily influenced by myths. In current debates, AI is often discussed from the perspective of whether it is a "hype." The term "hype," derived from the ancient Greek "hyperbole" (exaggeration), colloquially describes a short-lived trend created by the media. It often involves exaggerated attention that can be misleading and may quickly fade. The well-known Gartner Hype Cycle model by Jackie Fenn from the Gartner Group describes the phases of public attention towards technological innovations. According to the model, the cycle begins with a technological trigger that raises high hopes due to lofty expectations and performance promises. This leads to the "peak of inflated expectations," followed by a plunge into the "trough of disillusionment." Subsequently, the "slope of enlightenment" follows, ideally leading to a moderate and stable rise to the "plateau of productivity," where the technology finally finds practical and sustainable applications.

The hype corresponds to what was previously described as the first level of "myth." It questions the rational justifications for the possibilities and limitations of AI and separates media over-exaggeration from actual capabilities. The “real” myth of AI relates to the second level, concerning non-rational motivations. From the perspective of salvation and damnation, two main mythical motifs can be distinguished:

- Myth of the Thinking Machine: The notion that AI can perfectly simulate the cognitive abilities of the human mind is a cultural myth reinforced by analogies and future visions.
- Frankenstein Syndrome: The fear that AI will become super intelligent and take control over humans distracts from more realistic problems such as "artificial stupidity," where AI systems can thwart human goals.

These motifs can be readily identified as non-factual, mythical narratives when considering the limitations of AI. Artificial Intelligence operates based on data collected for specific purposes, which inherently lack objectivity. Different purposes would require different datasets, resulting in biases and discrimination intrinsic to the data used. Moreover, AI systems and the platforms, services, business models, and organizations built on them are developed by individuals with specific interests and values, further compromising the objectivity of AI. The frequently discussed issue of bias and lack of diversity is merely the most apparent symptom. Technological progress is deeply ambivalent and often leads to unintended and unforeseen consequences, such as increased CO_2-emissions. The idea of salvation through AI is not grounded in scientific research but rather resembles a quasi-religious belief. These themes will be further explored in the following chapters.

3. Generative Artificial Intelligence

Artificial Intelligence (AI) refers to the concept of implementing a behaviour of computer systems that resembles human intelligence. These processes include learning, reasoning and the ability to adapt.

In the initial phase of AI, research focused on symbolic AI, which aimed to encode human knowledge explicitly through logical rules and symbolic representations, much in the tradition of the classical programming approach for the engineering of computer systems. This approach led to the development of early AI systems capable of performing tasks such as theorem proving and game playing and led to promising results, but did not scale well. In the 1980s, a paradigm shift occurred with the advent of machine learning, a subfield of AI that emphasizes the development of algorithms enabling machines to be trained from data without explicit programming and make predictions based on the models generated from that data. The rise of the so-called neural networks, particularly deep learning around 2009, revolutionized AI by allowing models to automatically discover intricate patterns in large datasets. Once again scaling the models was tricky and took decades to be resolved benefitting from a) the availability of larger computational resources, b) the increasing amount of digital data and c) new algorithmic approaches (deep learning).

The concept of Generative AI (GenAI) refers to systems that can generate content mostly in the form of text, but also as images, music, movies, or even code in programming languages. Among the various architectures propelling this innovation, Large Language Models (LLMs) have become central, offering unprecedented capabilities in natural language understanding and generation. The availability of the application ChatGPT by OpenAI in 2022 put these technologies and generative AI in general into the awareness of ordinary users and has emerged as one of the most transformative technologies within the IT sector since then.

The origins of GenAI can be traced back to foundational models like Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks. These models were adept at handling sequential data and capturing temporal dependencies. Limitations in parallelization and handling long-term dependencies prompted the search for more robust architectures. A seminal moment in the history of generative AI was the introduction of the Transformer model (Vaswani et al. 2017). This architecture is utilizing a mechanism known as self-attention to process entire sequences of data simultaneously. This innovation allowed for greater parallelization, significantly enhancing computational efficiency and the capacity to model long-range dependencies in data.

Transformers paved the way for the development of foundation models in subsequent years. Foundation models are referring to large-scale models that pre-trained on vast datasets encompassing diverse domains, endowing them with a broad under-standing of language and world knowledge and that serve as a robust base for a variety of downstream tasks. Fine-tuning these pre-trained models on specific tasks or domains allows for the rapid development of highly specialized AI applications.

Among the first applications of the transformer architectures have been BERT and T5 both by Google. While BERT focuses on understanding the context of words in a sentence by considering both the left and right context simultaneously, T5 has been trained on a massive dataset of text and code using a text-to-text framework that enables it to perform the text-based tasks that they were pretrained for such as translation, restoring corrupted text and analysing grammar in sentences. The GPT family by OpenAI also apply the transformer technology, and version three becoming the backbone of the first version of the successful ChatGPT application. GPT and ChatGPT still stand as one of the most prominent LLMs as it can generate human-like text based on a given prompt. Its strengths lie in its versatility and fluency, capable of performing a wide array of tasks without task-specific training. How-ever, competing models such as Google's Gemini, Anthrophic's Claude and the open-source models Mistral and Llama (by Meta) are representing a strong competition and the technologies and the related products evolve at rapid speed. Chatbot Arena is an interesting resource to get an overview on the dynamic developments in the field of LLMs (Chiang et al. 2024).

As generative models generate data samples that resemble the training data as a general mode of operation, their internal model represents only an implicit representation of the common-sense world knowledge that was embedded in the source data. This leads to a black box system that cannot be understood in the same way as engineered systems could be studied and understood. This situation if amplified by the fact that the models are large making it impossible to be understood at the lowest level of detail. Secondly, due to the algorithmic determination to forecast the most likely next sequence of words that follow the words already generated, the models tend to "hallucinate" by generating wrong or misleading statements or data. Both facts combined create a potential problem in application and reliability of those systems, contributing to the fact that the systems are mystified in a sense – mystification is by design, one could say, due to the architecture. Those issues are being worked on under the context of explainable AI (xAI), which often also uses LLM to explain the output of another LLM – once again adding to the mystification of the overall system, if no or little human control is incorporated in the process. A detailed explanation of GenAI systems can be found in (Patil and Gudivada 2024).

4. Effects of the Mystification of AI

In the following section the effects of mystification are discussed in respect to AI, more specifically the novel forms of Generative AI (GenAI), that became prominent with the release of the first Large Language Model by OpenAI in 2022 and subsequent models by the company and its competitors. Several perspectives are taken, namely economics, education, society, transhumanity and superintelligence to point out the mystifying aspects and current and potential future im-pacts that they might have.

4.1 Economics

Companies worldwide invest billions of dollars in AI to optimize their businesses, reduce costs, and boost productivity. According to a report by Bloomberg Intelligence, "the generative AI market is poised to explode, growing to $1.3 trillion over the next 10 years from a market size of just $40 billion in 2022" (Bloomberg Intelligence 2023). However, these developments are ac-companied by myths and misconceptions.

One prominent misconception is that AI systems make autonomous decisions and act independently. This notion, often referred to as "The Myth of Technological Determinism," suggests that technology directs its own development and application without human intervention. Head-lines like "AI could replace the equivalent of 300 million jobs" give the impression that AI systems have their own initiative. In reality, it is always humans who decide how AI is used and what economic, social, and societal consequences these decisions have.

These misunderstandings lead to "sociotechnical blindness", where the human influence on these systems is overlooked or underestimated. AI systems do not act independently; rather, it is the developers, managers, and users who set their goals and parameters. By ignoring the human contribution and decision-making in technology development, we risk failing to recognize the responsibility for the consequences of these decisions. The process of finding an agreement or a position on how to steer and regulate the use of AI technologies is currently underway and emerging. It is interesting to note that those regulations initiatives are different in different countries such as Europe and in China and the United States, also representing different approaches on technology management in general.

4.2 Education

The education sector and is impacted heavily by the availability of GenAI technologies and applications, both on the student side and on the lecturer side. An immediate reaction on the use of GenAI applications for writing assignments by students was the regulation by education organizations, to ban the new technology. Meanwhile,

the reactions are more nuanced and the awareness is building up, that the new technology will not go away. The education process will change over the next years. There are positive and negative aspects to be discussed in turn.

On the positive side we see the advent of new tools that might help to accelerate the cumbersome steps in academic processes, e.g., a support in the research process by providing more sophisticated tools to find relevant research works, provides new ways of accessing them by using natural language-based interfaces and even multimodal access to relevant information sources. The writing process might be supported with tools to improve language and style of the written artifacts as well as writing out full paragraphs based on a set of instructions, keywords, and notes by prompting. These techniques might make research activities more accessible for those students that have an issue in writing up their research findings. While the tooling in GenAI might support academic work, there is still a debate how beneficial the acceleration or automation really is.

For the learning process on the student side there is a strong potential to support the individual learner with individualized content that is tailored to his or her needs, again being inquired using the language and style of the student and being able to take into account different levels. This aspect of explaining in a repeated way in a human fashion can support self-directed learning quite efficiently. It also addresses learning support in a scalable way as there is human bottleneck in terms of limited amount of time or availability of tutors. On the side of the lecturer, we do see the advantage of a faster creation of teaching materials with the help of GenAI as a part of the preparation or even during the lecture as a live teaching element. Further, the technologies can be used to generate candidates for questions for the students that they use to verify their learning progress.

GenAI technologies and applications are currently mostly available in a centralized way as web-applications or using programmable interfaces. This is due to the high demands on computational resources needed to build and run these models. Smaller and more affordable models will become available in the (near) future. This situation leads to a black-box-characteristic when being used by engineers to build custom application. The functionality of such a black box is by definition not known and thus becomes a myth in terms of functions and abilities. Which leads to a partly speculative behaviour of the overall system. For technical engineering professions, it becomes paramount to develop a good understanding of the inner workings of GenAI technology.

Finally, the motivation to learn a skill from the grounds up before using that skill with the help of GenAI will be a challenge in the education domain. It is important to convey to the coming generation of students, that it is very valuable to study a skill and to obtain a competence, before that skill is being augmented and amplified by the use of GenAI.

4.3 Society

Generative AI technologies have a high application potential in the generation of multimodal digital content in the form of texts, images and even short movies. This aspect relates to the media sector and is impacting the classical field of Journalism as well as the social media domain, challenging the importance of human authorship and the originality and trustworthiness of generated content. AI-Generated Content (AIGC) poses new challenges on the ethics of creating and consuming in the digital context and Deepfakes, Misinformation and Dis-information are posing an immediate thread on the way how the society perceives the dominant digital information sector. Due to the amount of generated digital content, we might create a digital reality that cannot be trusted anymore, since one cannot tell the difference between generated and original content and between fact and fiction. As a result, this might lead to a minimizing of trust in any form of digital content, putting more importance on the aspect of social interactions in the physical reality (“face-to-face”). While this form of interaction might have the advantage of better connecting and bonding (Böhm, Sammet, and Schmidt 2023) it also reduces communication productivity. We might not use digital communication for important settings and rely on physical meetings instead. Taken to the extreme, we might not even rely on written communication, as this could be generated as well. Which brings us back to the society of spoken language (only) and direct interaction with (known) peers, which would be a step back in the past. Therefore, it is quite significant to be able to maintain trustworthy digital relations despite the use of generative technologies, working against a mystification of the digital reality.

Related to the observations above is the impact on societal opinion building and the need to protect important societal processes like elections and other democratic decision processes. Generative AI can be used as an amplifier to generate false or misleading information at large scale to create strong opinion leaders that are not driven by facts, realities, and majorities but ideology.

Finally, and overuse of GenAI can also have a substantial impact for the social networks. Negative aspects being that the flooding with generated content might make it unusable (e.g., too much content to cope with) or not interesting (e.g., too generic content), rendering a social medium less useful. On the optimistic side there might be the option that relevant content might be created faster and in a more precise way (correct language, spelling, tone and length) and in a more intriguing way (by adding captivating illustrations, visual representations and the like), making it more attractive and even more accessible for content creators with an important message but not unlimited resources to create professional and compelling content otherwise.

4.4 Transhumanism and Superintelligence

Transhumanism and superintelligence are popular concepts that claim to offer realistic views on AI's potential, but are quickly revealed as myths upon closer examination.

The core idea of transhumanism aims to optimize the human body and mind through technological means. In transhumanist thinking, the human body is primarily viewed as a machine that can be improved by technology. The goal is to transcend the limitations of the human body and mind. The grand utopia of transhumanism is the conquest of death. Through the concept of "mind uploading," it is envisioned that consciousness can be stored in digital form, thereby achieving immortality. This idea strongly resembles religious notions of the afterlife. The crucial difference is that, in transhumanism, immortality is not granted by the grace of a transcendent God. Rather, transhumanism sees humans as fundamentally capable of redeeming themselves. Such self-redemption through technology has the character of a belief and is not the result of scientific research. This expectation of redemption is associated with a devaluation of the present, as it is something to be overcome. This attitude can lead to a certain passivity toward current challenges, since the "real" objective lies in the future—a motive that parallels the expectations of early Christian communities convinced of the imminent return of Jesus Christ.

The concept of an Artificial Superintelligence (ASI) also takes up the mythical motif of overcoming the inherently flawed human condition. It refers to the idea that it is possible to develop an artificial intelligence that significantly surpasses human cognitive abilities in all areas. This concept is often discussed in the context of the “technological singularity”, a hypothetical future point at which artificial intelligence exceeds human intelligence and possibly leads to an explosive improvement of its own capabilities. The term was coined by the mathematician and science fiction author Vernor Vinge, who predicted in 1993 that superintelligence would be achieved within 30 years. One of the most prominent advocates of the singularity is Ray Kurzweil, Director of Engineering at Google. He predicts that by 2029, artificial intelligence will reach human intelligence, referred to as Artificial General Intelligence (AGI) and surpass it by 2045 (ASI). However, the feasibility of superintelligence is highly contested. Various studies discuss the energetic, technological, and economic challenges involved. For instance, one study concludes that superintelligence would consume more energy than is available in industrialized countries. Additionally, the potential dangers posed by such superintelligence are also debated, such as the coexistence between human intelligence and the artificial superintelligence (Russell 2019).

Similar to transhumanism, the concept of superintelligence draws on familiar myths: predictions of the singularity create a constant sense of anticipation, which persists despite missed deadlines, fostering a sense of inevitability akin to the early Christian belief in the impending end times. The promise "The Singularity is near" echoes the eschatological formula of the end of the world and the arrival of the kingdom of God (Rayhan 2023).

5. Conclusion

Enlightenment thinkers have long known that education is the most important tool in demystifying myths. In "The Age of Technology Mysticism," comprehensive skill development appears to be the decisive point. Only through extensive knowledge and abilities can misunderstandings be clarified and technologies like AI be properly classified and evaluated. Education serves as a basis for knowledge that enables distinguishing between myth and reality. This requires skills, knowledge, and attitudes that allow people to handle AI critically.

Many organizations and orders of governance have accepted this need and have built elaborate frameworks of so-called “Future Skills” (Kotsiou et al. 2022). For enlightenment in the age of digital mysticism, three skills seem particularly appropriate, both for the system and application designers but also for the users of those applications:

- Knowledge of How GenAI-systems operate: To properly classify AI, a basic understanding of how Generative AI Systems function is essential. Without understanding how these models operate, their work remains opaque, leading to misplaced trust or mistrust and one cannot critically assess their procedures and outcomes.

- Dealing with Uncertainty: Despite comprehensive knowledge about GenAI and LLMs, it is (currently) impossible to explain how AI arrives at its conclusions. This "black-box phenomenon" makes it difficult to interpret and trust the results of LLMs (Von Eschenbach 2021). Current approaches of Explainable AI (xAI) are partial solutions at most as they counterfeit a black box with another black box explaining the first one. A competent approach to deal with this uncertainty is crucial.
- Critical Thinking: Critical thinking is indispensable for analysing and evaluating AI results. It is also essential for uncovering the narratives and myths surrounding AI. Historically, Socrates exposed myths through targeted questioning in his dialogues, and Kant urged people to use their own reason. This applies today as well: we must be brave enough to question the promises of AI and recognize the underlying economic and political interests.

Enlightenment in the "Age of Digital Mysticism" can be understood as the constantly ongoing project of protecting people from irrational fears and exaggerated hopes. Kant emphasized the importance of using one's own reason. Today, this means demystifying the myths of the digital era. Through education and scientific enlightenment, we can better understand the actual capabilities and limitations of AI technologies and make informed decisions. Its good or bad uses, can only be ensured if we have "the courage to use our own understanding", according to Emanuel Kant.

To conclude, it seems that technological complexity and the exponential development of new models, functions, and applications in the domain of Generative Artificial Intelligence currently increases the myths that are built and perceived around it leading to a phase of technology mysticism. Connecting to the history of human development we are optimistic that phase of technology enlightenment with respect to the emerging technologies lies ahead that will help us to understand and use the generative technologies in a responsible and beneficial way. As the past teaches us, this future will not happen automatically, but requires active shaping by all of us.

References

Bloomberg Intelligence. (2023). *Generative AI to Become a $1.3 Trillion Market by 2032*. Bloomberg L.P., 2023, sec. Announcements. https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/.

Blumenberg, H. (1985) *Work on Myth*. Studies in Contemporary German Social Thought. Cambridge, Mass.: MIT Press.

Böhm, K; Sammet J; Schmidt, J. (2023). *The Social Accumulator as a Concept to Manage Social Energy in the Age of Digital Transformation – An Explanation Model for Digital Interaction among Human Actors*, International Journal on Advances in Systems and Measurements, ISSN 1942-261x, vol. 16, no. 3 & 4, year 2023, http://www.iariajournals.org/systems_and_measurements/

Chiang, Wei-Lin, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, et al. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. arXiv. https://doi.org/10.48550/arXiv.2403.04132.

Kant, I. (1991). *Kant: Political Writings*: Edited by H. S. Reiss. Translated by H. B. Nisbet. 2nd ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511809620.

Mayor, A. (2018). *Gods and Robots*. 27 November 2018. https://press.princeton.edu/books/hardcover/9780691183510/gods-and-robots.

Patil, R.; Gudivada, V. (2024). *A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs)*. Applied Sciences 14 (5): 2074. https://doi.org/10.3390/app14052074.

Rayhan, S. (2023). *AI Superintelligence and Human Existence: A Comprehensive Analysis of Ethical, Societal, and Security Implications*. https://doi.org/10.13140/RG.2.2.12922.26561.

Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York, Viking.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, L.; Aidan J.; Gomez, N.; Kaiser, L.; Polosukhin, I. (2017) *Attention Is All You Need*. In Advances in Neural Information Processing Systems. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Von Eschenbach, Warren J. 2021. *Transparency and the Black Box Problem: Why We Do Not Trust AI*. Philosophy & Technology 34 (4): 1607–22. https://doi.org/10.1007/s13347-021-00477-0.

Artificial Intelligence and Trend Forecasting in the Fashion Industry: Ethical and Anticipated Ethical Issues

Olivia Bowles and Richard Wilson
Towson University, USA

obowles1@students.towson.edu
wilson@towson.edu

Abstract: Trend forecasting within the fashion industry is aimed at predicting the of future popular styles, materials, colors, and all things related to the development of fashion. Trend forecasting is based on information collected about past, present, and projected future developments in the fashion world. AI, as it has been applied in other areas in society, is now being applied within the fashion industry. This analysis will focus on AI being applied to trend forecasting within the fashion industry. Using AI in trend forecasting brings a number of advantages for those in the fashion industry while at the same time raising a number of ethical concerns. Three significant ethical concerns that are related to the employment of AI in the fashion industry are, the invasion of privacy from the data mining required to gather data to make predictions about fashion trends, the consequences for businesses of incorrect trend predictions, and the anticipated environmentally unsustainable nature of fast fashion due to the overconsumption that consumers practice and that is increasingly promoted by AI. Overconsumption is the product of how AI is involved in rapidly forecasting new trends in fashion that consumers then follow. Addressing all of these ethical issues is needed because fast fashion has continued to be demanded by both consumers and investors. New technologies, particularly AI, have allowed fashion brands to develop methods that allow them to be ahead of rapidly changing fashion trends and put high-demand products in stores faster, increasing the prevalence of fast fashion within the fashion industry. Questions now arise about how to address these ethical concerns. Important questions for the fashion industry include what issues will the use of AI in trend forecasting create in the long run within the fashion industry, and how can proactive action be taken within the fashion industry to prevent future ethical issues in the fashion industry? This paper which employs empirical methods to describe the fashion industry while also employing conceptual analysis related to philosophical ethics. The goal of the analysis is to provide, an informative ethical analysis of AI in trend forecasting, while also attempting to develop ethical guidance for concerns involving the use of AI in trend forecasting. The paper perform an anticipatory ethical analysis that attempts to address future concerns about fashion and will close by drawing conclusions about the direction of future analysis related to the application of AI in the fashion industry.

Keywords: Artificial intelligence, Trend forecasting, Fashion industry, Empirical and Conceptual analysis, Ethics, Anticipatory ethics

1. Introduction

The fashion industry is a broad and complex industry that features many different materials, styles, and methods of craftsmanship related to the development of fashion. Dillon (2018) states that prior to the mid-19th century, clothing was mostly custom-made by either individual tailors who designed, cut, fitted, and finished clothing or individual dressmakers who copied/ adapted the latest clothing ideas from Europe. Dillon (2018) goes on to state that in the 19th century, retail outlets developed, and clothing began to be mass-produced, as well as fixed in price and sizes. Dillion also states (2018) that at first there were three main categories in the fashion industry: haute couture fashion, ready-to-wear fashion, and mass-market fashion. Haute couture originated from dressmaking and the first couture house was established in the 1850's. Haute couture is still designed by fashion houses around the world and is considered the most prestigious kind of fashion because it is custom-made, extremely expensive, and exclusive. Haute Couture is so exclusive that it is in fact, "governed by Paris's Fédération de la Haute Couture et de la Mode" (Bala, 2019). Ready-to-wear fashion, as Dillion (2018) states, is shown in fashion shows, similar to haute couture, but is delivered to the marketplace months after it is designed and developed, and unlike haute couture is sold at a fixed price and in fixed sizes. Mass market fashion involves a larger scale production of fashion products, and due to cheaper methods of manufacturing, can be significantly cheaper than haute couture and produced in a way that makes this clothing, ready-to-wear. In addition to these types of fashion, in recent years a fourth category within the fashion industry has arisen, fast fashion. Fast fashion is similar to mass-market fashion but is arguably different due to its design, production, and market placement at speeds which are much faster than the other types of fashion. Due to the introduction of fast fashion into the fashion world there is a substantial difference between how earlier types of fashion were produced and how fast fashion is produced and how it influences the fashion world. This difference between types of fashions creates a gap between how fashion is made and the result of how fashion is consumed. Fast fashion has an environmental impact that poses a variety of problems for the general public but that are largely unrecognized by the fashion industry and members of the public. This

analysis will employ a conceptual ethical analysis to attempt to close this gap. The analysis will proceed by discussing using empirical analysis fast fashion, trend forecasting and how AI is employed with fashion. This discussion will be followed by a conceptual analysis of ethical and anticipated ethical issues with trend forecasting in the fashion industry.

In the marketplace of the fashion industry there are changes in fashion that are called 'trends' that appear. Fashion trends involve the spread of ideas and concepts about current fashions that appear in local, national and global arenas. These trends can be defined by attributes such as their color, shape, pattern, material, and the style of fashions. From haute couture designers, to streetwear, to fashions related to significant historical events, fashion trends can be set from and spread from a large range of places, people, and events. There are multiple theories that have developed about how trends work but a few that can be identified are the tickle-up theory, trickle-down theory, and the trend cycle. The Trickle-up theory of fashion occurs when trends originate from popularized fashions of everyday people and spread upwards to the domains of haute couture and high fashion. An example of trickle up fashion would be the popularization of streetwear. In the 1980's-1990's, streetwear, characterized by baggy clothes, ripped jeans, and sneakers were popularly worn by skaters and punk groups. These examples of a trend developing from streetwear caught the attention of many haute couture designers and examples of streetwear began appearing in haute couture and designer brands. Collaborations began to develop one of which appeared as the Louis Vuitton X Supreme 2017 fall collection. This collection combined the work of haute couture designer, Louis Vuitton, with popular streetwear designer brand, Supreme. In contrast with the trickle up theory in fashion is the trickle-down theory of fashion. Trickle-down theory occurs when trends originate from haute couture and high fashion and spread down to mainstream fashion in the mass market and fast fashion levels of the fashion world. Müller (2008) states that prominent haute couture designer, Christian Dior, had his first fashion show in 1947 which changed many women's desires in fashion from a wartime silhouette style to a retro-romantic style. The trend cycle in fashion refers to the representation of the life of a trend. What occurs in the trend cycle of fashion can be described as follows. First the trend is introduced, then it rises in popularity, then the trend reaches its peak at maturity until it declines in popularity, and then the trend is eventually rejected. The trend cycle can range in length depending on the nature of the trend, including the fact that some trends even reappear after they have been rejected.

2. State of the Art: Fast Fashion, Trend Forecasting, AI in Fashion and Stakeholders

2.1 Fast Fashion

Fast fashion is a more recent type of fashion that has prevalently emerged in recent decades where trends arise and then fall out of fashion just as quickly as they arose. Fast fashion brands and retailers are constantly producing and supplying the latest trendy apparel to the public at a rapid pace and lack the quality of items produced in other parts of the fashion industry. Even though much of the time fast fashion lacks in quality, consumers are still drawn to it because of its incredibly low prices, its ease of access and because it keeps up with the latest trends that people don't want to miss out on. This is captured in the phrase the fear of missing out (FOMO). "FOMO has a major influence on customer decision-making in both fast and slow fashion" (Bläse et al. 2023). FOMO when it occurs within fast fashion is problematic for a number of reasons. In both social and physical environments there are three major reasons for why fast fashion is problematic, the large consumption of resources needed to supply the growing demand, the cheap prices that come from the exploitation of people, and the massive amount of fashion waste and pollution that comes from production and overconsumption of fast fashion. Fashion waste is the accumulation of large quantities of apparel and clothing scraps discarded in landfills, dumping sites, or otherwise disposed of in unsustainable ways. Of the 12,970,000 tons of clothing and footwear generated in the USA in 2018, nearly 70% was placed in landfills (EPA, 2018). The European Union, United States, and China are the main exporters of tons of fast fashion apparel and materials; and due to the low quality of the materials used in fast fashion, are disposed of in large amounts every week to countries, far away from the EU, USA, and China, in places like Chile and Ghana (Ziegler and Elpídio da Silva, 2022). Fashion brands and retailers may not have the intention of contributing to fast fashion and to fast fashion waste, but the desires for economic success has an influence upon them and guides them to participate in unsustainable and unethical manufacturing and production practices. Consumers aren't blameless either, their desire to participate in the purchase of new trends in fashion drives them to buy unnecessary amounts of clothing and their particular lack of desire to deal with unwanted clothing properly, contributes to fashion waste and overconsumption as well.

2.2 Trend Forecasting

Trend forecasting in the fashion industry is the effort to predict future popular styles, materials, colors, etc. in the fashion industry marketplace based on information collected about past and present fashion and based upon previous fashion trends. This data is used in conjunction with AI to project future fashion trends based upon what has previously occurred in the fashion industry. Forecasting provides brands and retailers with useful information related to understanding their audience, what products need to be restocked or marked down in price due to customer preferences, and guidance as to what types of features should be included in designing new products. Traditionally, fashion forecasters travel and observe global and local fashion developments, attend fashion shows and read fashion magazines, while studying social and economic standings of consumers, cultures, weather, political climates, and sales data since all of these factors are critical for predicting future fashion trends. With this information, they use theories like trickle-up theory, trickle-down theory, and the trend cycle analysis to attempt to understand how a trend will behave and to predict new trends. The processes related to fashion forecasting can be time-consuming, which is why it is not surprising that AI is being adopted into the fashion trend forecasting industry. The use of AI in the fashion industry reduces the time needed to engage in trend forecasting.

2.3 AI in Fashion

Artificial Intelligence (AI) since its inception has sought to enable machines that exhibit human-like intelligence and to perform tasks that resemble human reasoning processes. AI has become an integral part of many items that are used in day to day activities within society, and one of the areas where AI is now having an influence is in the fashion world. AI is used in fashion for shopping with style assistants, chatbots, and smart mirrors, designing products, improving manufacturing, developing demand and trend forecasting. In fashion trend forecasting AI can be used in the following ways: for trend forecasting in fashion, as a tool to collect large amounts of data, identify and label fashion attributes of images, and to identify and predict fashion trends based upon what has been identified in extremely large sets of customer data. Data can be collected through tools such as application programming interfaces (API's), which as Luce (2019) states, act as a communication layer between two programs enabling access and data collection on websites such as social media and shopping sites. With access to API's, natural language processing (NLP's) can be employed for a wide variety of purposes. For example, Luce (2019) states that API's can be used to gather and analyse text sentiment data from shopping websites and social media platforms, where customers discuss fashion products. NLP's can also be used to understand and analyse data collected through surveys. Luce (2019) also states that computer vision can extract features using edge detection and object detection. Additionally, Zhao et al. (2024) identity's how AI employing deep learning and deep neural networks such as region convolutional neural networks (RCNN), AI can identify and categorize objects within images. Zhao et al. (2024) identify that to develop this type of detailed analysis machine learning needs to be fed data from a large dataset to train the model to recognize attributes. Luce (2019) states that this can be done with supervised learning, where the model is trained on labelled data. Zhao et al. (2024) use unsupervised learning, where the model or neural network is trained on unlabelled data and receives validation data to learn. Additionally, Shi et al. (2021) state that image attributes can be labelled by using a deep learning method called transfer learning. Additionally, Shi et al. (2021) state that transfer learning, when applied to fashion images, can recognize clothing attributes based on the examination and interpretation of other fashion images. Leeway Hertz, an AI Development Company states, that a program can identify and predict fashion trends in a set of data by using machine learning methods and deep neural networks, which can learn unknown patterns, which in turn allows machine learning algorithms to draw conclusions and make predictions about the unrecognized patterns that occurred within the data. Zhao et al. (2024) state that in their trend analysis model, it is necessary to have an accurate identification and classification of images, which serves as the foundation for forecasting fashion trends based on the images that have been analysed. Zhao et al. (2024) claim that they calculate what a trend will be by looking at the frequency in which attributes and combinations of attributes appear in the images classified through the use of machine learning applied to images.

2.4 Stakeholders in Fashion Trend Forecasting

The primary stakeholders involved in the use of AI in fashion trend forecasting examined in this analysis include consumers, the fashion forecasting companies using AI, fashion brands and retailers using the AI fashion trend forecast results. Each one of these stakeholders is a primary stakeholder because they arguably hold the most influence or are most influenced by the use of modelling in AI infused fashion trend forecasting. The fashion forecasting companies using AI are important because they are the ones who are actively creating

the AI infused fashion trend forecast. Fashion brands and retailers are important because they are the ones using the AI infused fashion trend forecast to develop and manufacture fashions for the marketplace. Consumers are important because they are the target audience of the AI infused fashion trend forecast. It is also important to remember the three primary ethical concerns with trend forecasting in the fashion world, invasive data collection, consequences of incorrect trend prediction, and fast fashion contribution. To analyse these issues and attempt to provide guidance about these issues for anyone analysing trend forecasting in the fashion world standard ethical principles from Philosophy and the Association for Computing Machinery (ACM) Code of Ethics will be applied.

3. Ethical Analysis of Fashion Trend Forecasting

Fashion trend forecasting with AI involves invasive data collection, which brings up the ethical issue of the invasion of privacy. To produce more accurate forecasts requires the collection of more images and sentiment data by those employing the AI from consumers. This data can be gathered from social media, catwalk images, search history, and purchase history. Trend forecasting companies using AI like Worth Global Style Network (WGSN) and Heuritech use social media images for analysis. WSGN states that they use “image recognition technology to classify images” (WSGN) of their “proprietary global map of Instagram influencers” (WSGN). Heuritech states that due to “millions of images analysed from social media, computer vision and artificial intelligence” (Heuritech) they can accurately predict fashion trends. Zhao et al. (2024) states that a study by Gabale and Subramanian (2018) used an AI model that is trained on images from Facebook and Instagram to identify fashion trends. Depending on who is collecting this data and how this data is being collected, the data collection may be unknown to individuals whose data is being collected. If this data tracking was known to be occurring, it might be unwelcomed. The tracking occurs through processes such as social media tracking and through the collection of search and purchase history. From the perspective of consumers one ethical principle that applies to this situation, is Kant’s Categorical Imperative II. This ethical principles states that we should “Act so that you use humanity, as much in your own person as in the person of every other, always at the same time as an end and never merely as a means” (Ciulla, Martin, and Solomon 2013). This means that a person should never be used as a means to benefit yourself or others. Instead, the well-being of others, and their autonomy should be the end goal of adhering to the principle. This applies to trend forecasting agencies engaged in using AI are using consumers information and privacy as a means to get to an end of a more accurate trend forecast and ultimately as a source for their economic profit. In addition, section 1.1 of the ACM code states that we should strive to “contribute to society and to human well-being, acknowledging that all people are stakeholders in computing” (ACM, 2018). Collecting this information about consumers should not be strictly reliant on these more invasive methods. There are other methods like surveying that are an option for gathering data. WGSN has successfully implemented surveying into its trend forecasting process. There are also other routes to continue using social media which some forecasting companies may already perform, such as collecting data on a more transparent level by making social media users, customers, and browsers more aware of the data that is being collected from them, which would allow users the option of opting out of data collection.

From the perspective of trend forecasting agencies when trend forecasting is employed within the fashion world there is the possibility of failed forecasts. The implementation of AI into trend forecasting does not make this possibility disappear, but it does change the way failed forecasts can occur and introduces potential new risks that the use AI brings to the table. Failed forecasts can happen due to flaws in the AI as well as limitations of the AI compared to the abilities of humans. AI can have issues related to errors in coding or in the software program, there can be incorrectly labelled images, and there can be a lack of data on influential unaccounted-for events. The consequences of failed forecasts can include loss of profits, fashion waste, and waste of resources. These ethical issue conflict with another ethical issue, the invasive collection of data. They are related because more accurate trend forecasts come from an increase in the amount of data collected, which compels trend forecasting agencies to desire to use more invasive AI data collection for trend forecasts even though their intention may not be to invade consumers privacy. The principle of care ethics in particular applies as it is concerned with prioritizing of interpersonal relationships and individuals’ dependence on one another, highlighting compassionate and context-sensitive responses to other individuals and situations over rigid adherence to universal principles (Maio, 2017). A key component of care ethics is human connection and dependence on each other, this is something trend forecasting companies using AI should be able to implement. It is important that humans are involved in correcting and preventing mistakes related to the use of AI in trend forecasting. Balancing interests related to employing AI as a tool and not completely replacing human input with AI, is necessary. Care ethic’s flexibility and nonuniversal principles allows AI to be a tool and

it can include human input, while turn reducing AI's errors. Section 3.7 of the ACM Code of Ethics states that it is important to "recognize and take special care of systems that become integrated into the infrastructure of society" (ACM, 2018). As AI is integrated into trend forecasting in the fashion industry it is important to recognize that while AI is a sophisticated technology, it does have flaws and has the potential to be at times, wrong. AI when it used for trend forecasting has the potential to predict trends that will end up being incorrect. Due to unpredictable circumstances, trends could be predicted that lead to the production of products that won't be sold and that will end up becoming waste. For example, the AI forecasting company Heuritech, recognized Covid as a significant factor that would change their original predictions for summer and fall fashions in 2020 and decided to develop trend predictions with Covid in mind. An AI program cannot, at least, as of yet, fully take into account all influences that may affect future fashion trends. This responsibility is left to humans which is why it is crucial to not rely completely on AI and recognize that human input is still a necessary ingredient within the field of fashion trend forecasting.

The fashion industry is growing increasingly dependent on unethical, unsustainable practices which is largely a product of fast fashion. An important contributor to fast fashion is the increasing speed at which trends in fashion are changing. When fast fashion is conjoined with AI, following and creating new fashion trends has become quicker and easier to accomplish. AI makes it easier to gather, analyse, and utilize information and data for trend predictions and can make fashion predictions on its own. Some believe that AI trend forecasting can help with sustainability, by cutting down on the production of wasted "untrendy" products that people will not buy. Brad Lacey, the global design director of lifestyle footwear at New Balance, in an interview with AI trend forecasting company, Heuritech, stated, "people don't really think of efficiency as a way to be more sustainable but it's a big part of it, you know you look at how much waste is in fashion in general largely because we don't understand consumers desires and needs to the degree that we need to… understanding people's needs is a big part of sustainability" (Brad Lacey). While brands and retailers can use AI trend forecasting to aid in reducing the amount of unsold clothing by providing the products their customers want, it can also be abused as well. The overuse of AI trend forecasting in the fashion industry looks like it can lead to creating overflow in the market by miss predicting trendy products. Examples of overflowing the market with trendy apparel are companies like Shein, a dominant fast fashion brand. A better understanding of what the consumer wants will likely decrease the amount of unsold clothing waste, but that doesn't really help when other AI trend forecasting companies, brands, and retailers will abuse AI trend forecasting information to mass produce trendy items, likely in poor condition, and because of this, many of the products being produced will end up in landfills. This isn't just true with poor quality apparel items it is also true of decent or sometimes high-quality apparel. Consumers have the pattern of buying an item to wear once or twice and then end up discarding it. Or even if they donate it, Press (2016) states not all donations go to loving homes, most donated clothing is sold in bales that are sent to Eastern Europe or Africa for resale or end up as waste. Press (2016) states recycling apparel isn't of much help either as it is hard to recycle blended fabrics. Act consequentialism or act utilitarianism states that "An act X, is morally permissible if the consequences produced by doing X result in the greatest good for the greatest number of persons affected by act X" (Tavani, 2013). Examining the consequences of using AI in excess in trend forecasting leads to a number of consequences. For the stakeholders of trend forecasting, companies using AI consequences would likely benefit and profit from selling trend information. For fashion brands and retailer's, consequences would be increased sales and revenue. Consumers would receive the benefit of owning trendy apparel items. These are the more positive consequences, more notable and more concerning are the potential negative consequences which can be seen when observing the situation more carefully. Looking at the long-term consequences of the excessive use of AI in trend forecasting raises the question of what will happen to all of the trendy new clothing consumers buy? Trends have been changing more frequently with the development of technology and AI and brands are more able to predict these changes in trends and provide more trendy products, influencing consumers to purchase more frequently which is a major contribution to the ethical issues related to fast fashion. Fast fashion is a more pressing concern as it affects everyone on the planet as well as the ecosystem of the planet itself. This main point here is that since using AI in trend forecasting has such a large negative effect compared to the smaller more positive ones, under act consequentialism this excessive use of AI in trend forecasting would not be moral. This highlights why it is important to reduce consumption by consumers and to also reduce production on the part of the producers which includes AI trend forecasting companies, fashion brands, and retailers. Section 2.7 of the ACM code states to "foster public awareness and understanding of computing, related technologies, and their consequences" (ACM, 2018). Trend forecasting companies should make it apparent not only to their clients, fashion brands and retailers, but also to themselves that using AI in fashion trend forecasting is something to be used in moderation, and when used in excess contributes to fast fashion

and its negative effects because of trendy apparels attractiveness to consumers. Making fashion brands and retailers aware of these developments is important as they are the direct channels for consumers to consume apparel produced in the fast fashion industry.

4. Anticipatory Ethics and the Fashion Industry

With the rapid expansion of technology and particularly AI technology, it is important to recognize that the full effects of technologies has yet to be explored. Anticipatory ethics is a method for attempting to address the development of these new technologies and the consequences of this development. Anticipatory ethics is "focusing on the problems that can be anticipated as potentially arising because of emerging technologies" (Nestor and Wilson, 2022). To use anticipatory ethics, one can employ what are known as The Rules. The Rules are 5 guidelines that act as a way of attempting to develop a means for governance related to the moral responsibility of engaging with computing artifacts. The application of AI to the fashion industry involves the use of computing artifacts. The goal of The Rules is to "reaffirm the importance of moral responsibility for these artifacts, and to encourage individuals and institutions to carefully examine their own responsibilities with respect to computing artifacts" (Miller, 2011). Miller (2011) states that computer artifacts can be characterized as any artifact that includes executing a computer program. For the computing artifact, AI trend forecasting, rule number three will be used which is "People who knowingly use a particular computing artifact are morally responsible for that use" (Miller, 2011). Since fashion brands and retailers are the users of AI trend forecasting, which involves the use of computing artifacts, they are responsible for the anticipated outcomes of the use of those computing artifacts. Fashion brands and retailers are the providers of the apparel that consumer's desire. Fast fashion is on the rise, and product lifecycles are having shortened lifespan due to fashion trends emerging from fast fashion. Adding trendy apparel to the market with increasing rapidity due to the use of AI for trend forecasting will not slow fast fashion down, it will only make it faster. This will in turn worsen the effects of fast fashion. The relationship of the lifespan of trends and fast fashion becomes a key point of influence for the AI computing artifact. This exhibits why it is important that fashion brands and retailers use the computing artifact (AI) with caution in the fashion industry.

5. Conclusions

The application of AI in trend forecasting in the fashion industry is part of a significant shift occurring within the fashion industry which is also is moving forward with the introduction of AI in general to the fashion industry. AI fashion trend forecasting provides new opportunities for the fashion industry but also introduces new ethical concerns. As the fashion industry continues to adopt AI, stakeholders must recognize their responsibility of addressing ethical concerns that are arising associated with trend forecasting in the fashion world. These issues include privacy concerns, the consequences of inaccurate predictions, and the consequences of the introduction of fast fashion within the fashion industry. Trend forecasting companies can become more proactive by making their data collection more transparent and by empowering consumers with choices about data collection. They can attempt to make their clients, fashion brands and retailers more aware of the consequences of incorrect trend prediction and the contribution of fast fashion to environmental concerns. Fashion brands and retailers can moderate or limit the amount of AI trend forecasting they use and mitigate the contribution of AI to fast fashion by making less trendy clothing that is part of the fast fashion industry. Consumer habits develop in the current fashion marketplace, and they drive the trend cycle to move rapidly. Consumers buy into fast fashion and the mass production of trendy apparel that retailers and brands provide for consumers through the use AI for trend forecasting. The balance of using AI in trend forecasting within the fashion industry is not a one-dimensional problem that can be easily fixed, consumers need to know how AI will affect the fashion trend forecasting industry, what role they play in it, and how they can aid in the endeavour to address and treat the ethical concerns related to trend forecasting. Consumers can pay more attention to and place more mindful consideration on what they buy, thinking about the things they need versus the things they want or buy just to keep up with trends. By using anticipatory ethics, the fashion industry can proactively identify and mitigate potential negative consequences, creating a future where technological advancements enhance not only business success but also the well-being of all stakeholders involved. These goals can be accomplished when future research is directed towards fashion analysis and trend forecasting where additional subjects are included in our anticipatory ethical analysis. These subjects can include research on the role of virtual style assistants, smart mirrors and generative AI models employed by the fashion industry in the effort to develop more precise trend forecasting methods.

References

Association for Computing Machinery (2018). *ACM Code of Ethics and Professional Conduct*. [online] Association for Computing Machinery. Available at: https://www.acm.org/code-of-ethics [Accessed Aug. 2024]

Bläse, R., Filser, M., Kraus, S., Puumalainen, K. and Moog, P. (2023). 'Non-Sustainable Buying Behavior: How the Fear of Missing Out Drives Purchase Intentions in the Fast Fashion Industry'. Wiley Online Library, [online] 33(2). doi:https://doi.org/10.1002/bse.3509 [Accessed Jun. 2024].

Ciulla, J.B., Martin, C.W. and Solomon, R.C. (2013). 3rd edn. *Honest Work: A Business Ethics Reader*. New York; Oxford: Oxford University Press.

Dillon, S. (2018). *The Fundamentals of Fashion Management*. 2nd edn. New York: Bloomsbury Visual Arts.

Vijay Gabale and Subramanian, A. (2018). 'How To Extract Fashion Trends From Social Media? A Robust Object Detector With Support For Unsupervised Learning'.

Heuritech. (2024). *Trend Forecasting with AI: Fashion's Way Forward*. [online] Available at: https://heuritech.com/trend-forecasting-fashion-ai/ [Accessed Aug. 2024].

Heuritech. (2024). *Our success stories* - Heuritech. [online] Available at: https://heuritech.com/our-success-stories/ [Accessed Aug. 2024].

Luce, L. (2018). *Artificial Intelligence for Fashion: How AI is Revolutionizing the Fashion Industry*. New York: Apress Media.

Maio, G. (2017). Fundamentals of an Ethics of Care. *Care in Healthcare*, pp.51–63. doi:https://doi.org/10.1007/978-3-319-61291-1_4.

Miller, K.W. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. IT Professional, 13(3), pp.57–59. doi:https://doi.org/10.1109/mitp.2011.46.

Müller, F. (2008). *Fashion Game Book: A World History of 20th Century Fashion*. New York: Assouline.

Press, C. (2018). Wardrobe Crisis: *How We Went from Sunday Best to Fast Fashion*. New York: Skyhorse Publishing.

Nestor, Michael and Richard L. Wilson (2022). *Anticipatory Ethics and the use of CRISPR in Humans.* Springer Publishing.

Shi, M., Chussid, C., Yang, P., Jia, M., Dyk Lewis, V. and Cao, W. (2021). 'The Exploration of Artificial Intelligence Application in Fashion Trend Forecasting'. Textile Research Journal, 91(19-20), p.004051752110062. doi:https://doi.org/10.1177/00405175211006212.

Tavani, H.T. (2009). *Ethics And Technology: Ethical Issues in an Age of Information and Communication Technology*. Hoboken, N.J.: Wiley

US EPA. (2017). *Nondurable Goods: Product-Specific Data*. [online] US EPA. Available at: https://www.epa.gov/facts-and-figures-about-materials-waste-and-recycling/nondurable-goods-product-specific-data#ClothingandFootwear [Accessed Aug. 2024].

WGSN. (n.d.). *Trend Forecasting with Data Analysis and* AI. [online] Available at: https://www.wgsn.com/en/wgsn/data [Accessed Aug. 2024].

Zhao, L., Li, M. and Sun, P. (2021). 'Neo-Fashion: A Data-Driven Fashion Trend Forecasting System Using Catwalk Analysis'. Clothing and Textiles Research Journal, 42(1), p.0887302X2110042. doi:https://doi.org/10.1177/0887302x211004299.

Ziegler, A. and da Silva, A. (2022). 'The Dumping Grounds for Fast Fashion Clothes'. Latin American Journal of European Studies, 2(2). doi:https://doi.org/10.51799/2763-8685v2n2001.

Advancing Corporate Social Responsibility in AI-Driven Human Resources Management: A Maturity Model Approach

Marta Bubicz and Marcos Ferasso
Lusofona University, Intrepid Lab, Lisbon, Portugal
CETRAD, Porto, Portugal

p7052@ulusofona.pt
p7541@ulusofona.pt

Abstract: Artificial Intelligence (AI) in the corporate environment has been the subject of several current social debates of many scientific studies in reshaping human resource management (HRM) practices. There is a significant research gap in understanding how artificial intelligence (AI) can be ethically and effectively integrated into human resources management (HRM), particularly in relation to corporate social responsibility (CSR). This study aims to address this gap by proposing a maturity model to assess and guide the responsible implementation of AI in HRM practices. The research goal focuses on how AI can be aligned with CSR principles to ensure ethical, transparent, and socially responsible usage in organizational settings. The methodology includes a comprehensive review of 52 academic papers, employing bibliometrics, network analysis, and thematic content analysis to explore the interplay between AI, CSR, and HRM. These analyses allowed the identification of key ethical concerns and challenges in AI-driven HRM practices, such as bias in AI algorithms, data privacy issues, transparency, and the need for technical proficiency among HR professionals. The findings reveal a five-level AI maturity model, each stage representing progressive alignment with CSR principles. Organizations at lower maturity levels tend to have ad-hoc AI implementations with minimal CSR focus, while those at higher levels demonstrate full integration of ethical AI practices. Additionally, the study highlights the importance of transparency, accountability, and employee empowerment as critical elements for advancing AI maturity in HRM. This research contributes by offering organizations a practical tool to assess and enhance their AI-driven HRM processes through a CSR lens. It also provides a foundation for future research on strategic policy development, ethical AI governance, and continuous improvement in the integration of AI in HRM. Scholars are encouraged to explore these areas further, particularly in understanding how AI can foster not only organizational efficiency but also social responsibility and ethical standards.

Keywords: Artificial intelligence, Maturity model, Human resource management, Corporate social responsibility, Corporate governance, Technology management

1. Introduction

The rapid advancement of artificial intelligence (AI) has significantly transformed the landscape of human resources management (HRM), offering innovative solutions for recruitment, talent management, performance evaluation, and employee engagement. Companies across various industries are increasingly leveraging AI technologies such as machine learning algorithms, natural language processing, and predictive analytics to streamline HRM processes and enhance decision-making (Atanasov et al., 2023). However, the integration of AI in HRM also raises critical ethical and social responsibility concerns, including bias in AI algorithms, data privacy issues, transparency, and the need for technical proficiency among HR professionals (Grishunin et al., 2022; Chang and Ke, 2024; Struensee, 2021).

Recent studies indicate that AI use aids corporate sustainable development since it reduces costs and streamlines processes, generating greater productivity and resource optimization (Lega et al., 2013; Chen et al., 2024). Among the various applications, current research also highlights the increasing importance of the responsible implementation of AI tools, particularly in HRM (Akbarighatar et al., 2023; Chang and Ke, 2024). There is a growing concern that there may be a great dependence on AI in decision-making, which may bring personal harm, such as the degradation of professional capacity in future decision-making (Aitken et al., 2021; Londoño-Cardozo and Pérez de Paz, 2021). This concern is related to the moral and ethical aspects inherent to Corporate Social Responsibility (CSR) actions, but it brings to light the need for regulations in digitalization practices and the implementation of the use of AI tools by companies, defining the guiding principles of Corporate Digital Responsibility (CDR) (D'Cruz et al., 2022; Chang and Ke, 2024; Londoño-Cardozo and Pérez de Paz, 2021; Damoah et al., 2021; Krkač, 2019).

Given these complexities, it is imperative for organizations to not only adopt AI-driven HRM practices but ensure the alignment with CSR principles of positive social impact (Keenan et al., 2019; Atanasov et al., 2023), as well as CDR (D'Cruz et al., 2022; Chang and Ke, 2024; Londoño-Cardozo and Pérez de Paz, 2021; Damoah et al., 2021; Krkač, 2019). Identifying the stage of maturity of AI use in HRM involves evaluating how companies deploy AI technologies, the purposes they serve, the extent of their responsibility and ethical considerations, and the technical skills of their workforce. This is a practical limitation and a theoretical gap to be explored. Thus, this

paper introduces a maturity model approach to assess and enhance CSR practices in AI-driven HRM. The proposed maturity model provides a structured framework to evaluate an organization's current state of AI integration in HRM and to guide the development of more responsible and ethical AI practices within the organization's context.

This paper is structured to first review the existing literature on CSR in HRM and the current state of AI integration in HRM processes. Then, it introduces the maturity model, detailing the characteristics and assessment criteria for each proposed level. Finally, the paper offers practical recommendations and implications for theory and practice. Through this comprehensive approach, the paper aims to provide a valuable framework for organizations seeking to navigate the evolving landscape of AI in HRM responsibly.

2. Research Design

This literature review departed from the combination of bibliometrics, network analysis, and thematic analysis following the procedures implemented by Ferasso, Tortato and Ikram (2023), Beliaeva et al. (2022), and Ferasso et al. (2020).

2.1 Data Collection

Scopus was selected as the preferred scientific database due to the comprehensive collection of records it possesses related to the Business Management literature (Beliaeva et al., 2022; Ferasso et al., 2020). The Boolean search used the following searching terms: (artificial AND intelligence OR ai) AND (corporate AND social AND responsibility). We restricted the search to the Business, Management, and accounting area, and also limited to article and review documents. Additionally, we selected only the papers published in journals and in English language. No time frame was restricted in order to collect the comprehensive field on the edge of AI and CSR fields. These procedures reported a total of 52 papers considered for analysis. The *.CSV and *.RIS files were stored, containing the metadata from the 52 papers, to proceed with the network analysis procedures.

The bibliometrics used in this research considered the growth of publications per year, the most productive countries, and the main journals that published the sampled papers per year of publication.

2.2 Network Analysis Procedures

For the network analysis, the CiteSpace software was selected due to the comprehensive analysis it provides (Ferasso, Tortato & Ikram, 2023; Beliaeva et al., 2022). After converting the references from the sampled documents, the CiteSpace recognized 3,481 references from which 3,363 (96.6%) were considered valid for calculations and forming the networks. This percentage is in accordance with the valid range references to create a network.

Citespace was programmed to screen the dataset using title, abstract, and author keywords (DE) text processing; the node types considered references and keywords; and the selection criteria considered 100% of the most cited/occurred items from each slice (Top N%). For the Top N, it was set up to 100 levels of most cited/occurred items from each slice, while the time slicing of one year per slice recognized references expanding from January 2015 up to December 2024. Additionally, the following properties were selected for the CiteSpace project: LFR: -1.0; LBY: -1.0; L/N: -1.0; percentage of nodes to label (%): 100.0; TopN={n|f(n)≥e}: 2.0.

2.2.1 Thematic content analysis

After forming the networks, the most addressed topics were inspected according to the related literature, and the contributions were selected for building the AI maturity progression framework (Beliaeva et al., 2022; Ferasso et al., 2020). The results of thematic content analysis are shown in Table 1.

3. Findings and Discussion

3.1 Descriptive Statistics

The following results present the descriptive statistics taken from Scopus metadata information available in the said scientific database.

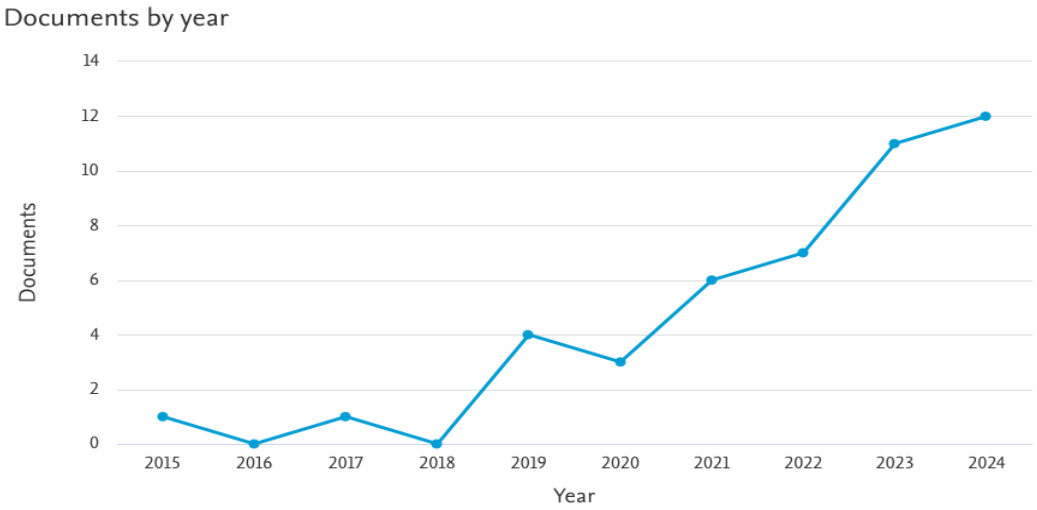


Figure 1: Publications evolution by year of paper published (Source: Scopus metadata)

The results shown in Figure 1 evidence that the interest started in 2015, but it was in 2019, with the publication of 4 documents, that the field grew. Since 2020, the total number of papers published in a constant way, evidencing that the field is gaining momentum. Despite the growing number of papers, the maximum number of papers published in 2024 was only 12 papers, which is explained by the expansion of AI studies in the Business Management literature. Therefore, the field of AI and CSR proved to be an interesting research area to be covered in future studies.

The most productive countries (Figure 2) addressing AI and CSR are the United States, with 12 documents published, followed by the United Kingdom with 8, and China with 7. These countries are on the edge of the North America-Europe-Asia axis of scientific production. Germany stands out with 4 publications. Australia, Canada, France, India, and Taiwan are the remaining countries with 3 publications each, and the Netherlands with only 2. Despite the top 3 countries belonging to the North America-Europe-Asia axis, it is interesting to note that studies are popping up from different countries worldwide, which underlines the curiosity of international scholars.

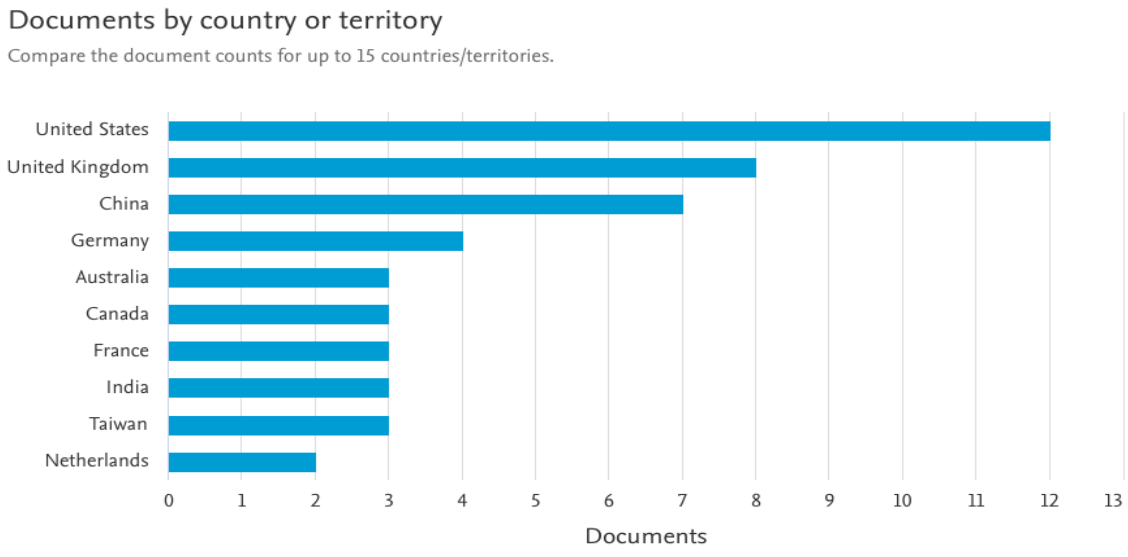


Figure 2: Most productive countries by number of publications (Source: Scopus metadata.)

Inspecting the top 5 journals (Figure 3) that published papers on AI and CSR, it is curious to notice that the Journal of Cleaner Production published two papers in 2021 on these topics, while the Journal of Business Ethics published the other two papers in 2023. All the other journals (Intelligent Systems in Accounting, Finance and

Management; Technology in Society; Asia Pacific Journal of Marketing and Logistics) published only one paper per year in selected years. These results prove that the AI and CSR field is at its birth stage.

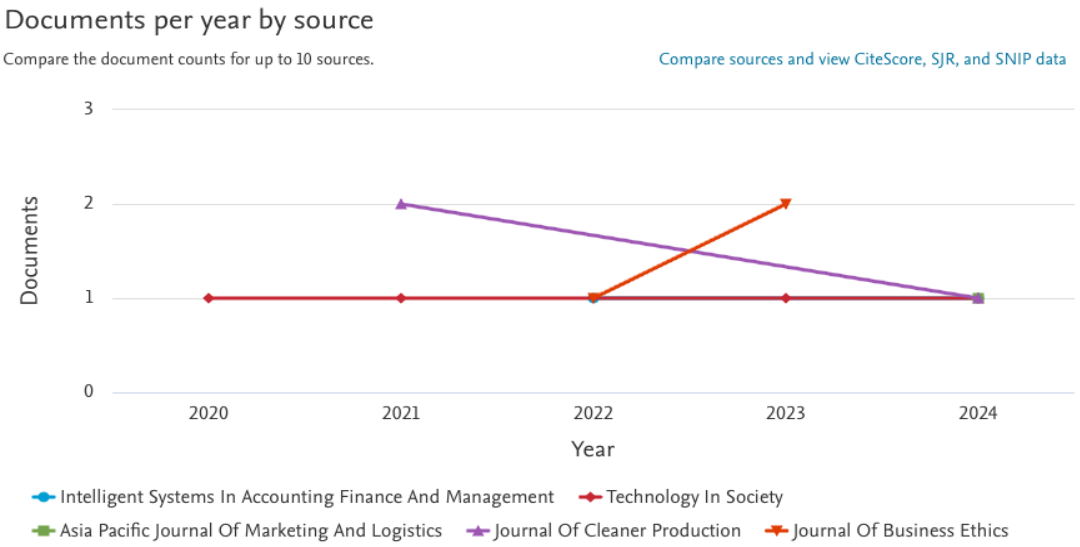


Figure 3: Published papers per academic journal per year (Source: Scopus metadata)

3.2 Network Analysis

Three different data analyses were chosen to form the networks for this study. The first is the co-citation network, in which the references are shown according to the citing and cited references formed in a comprehensive network. It is possible to notice the most cited references in the sampled documents, as shown in Figure 4. The thematic clusters were also formed based on the co-citation network (Figure 5), evidencing the grouped references according to specific topics that served as the basis for the thematic analysis. Lastly, the timeline view presents the same network in different layers formed by the identified clusters and distributes the references and keywords in one graph (Figure 6), which allows the identification of relationships not only among the references but also among references and keywords.

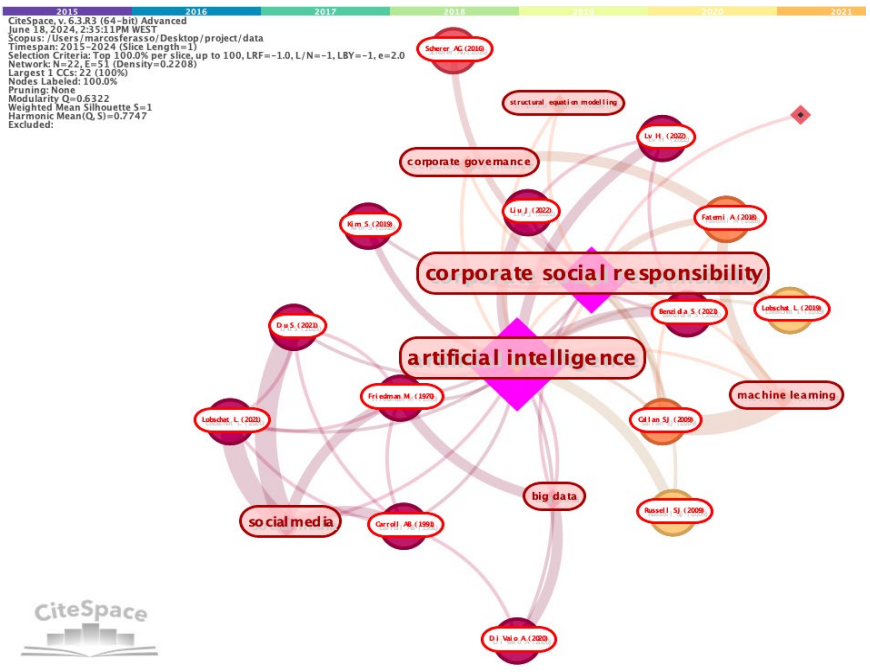


Figure 4: Co-citation networks and keyword network from sampled documents (Source: Scopus metadata)

Interestingly, machine learning, social media, and corporate governance are the keywords identified in association with AI and CSR.

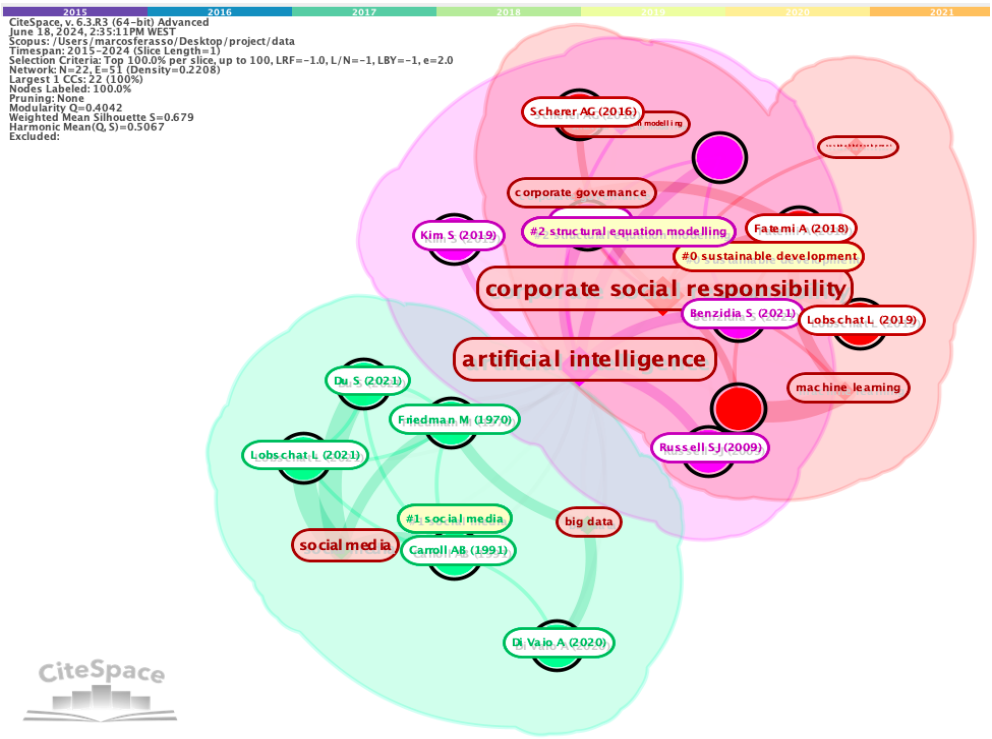


Figure 5: Thematic clusters of references and keywords from sampled documents (Source: Scopus metadata)

The findings allowed the identification of three thematic clusters of well-defined groups of scholarly contributions at the edges of AI and CSR.

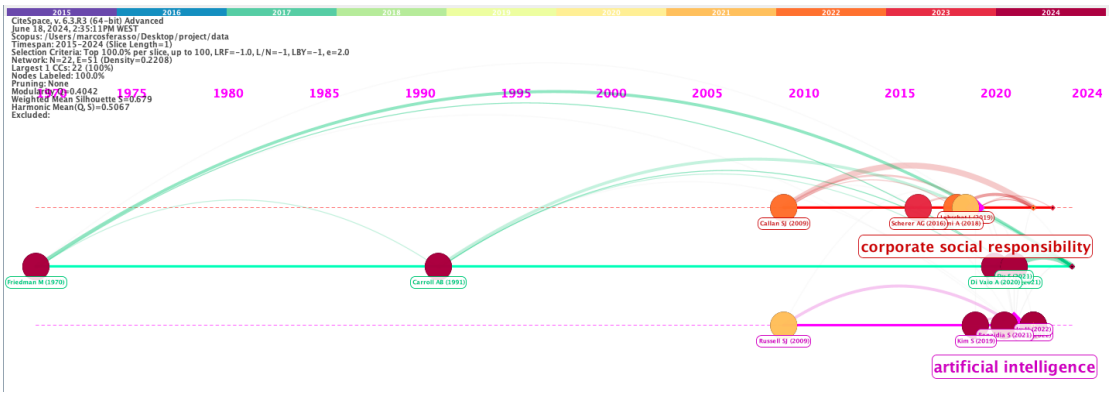


Figure 6: Timeline view of references and keywords, per year and per cluster, from sampled documents (Source: Scopus metadata)

3.3 Thematic Analysis

After identifying the key literature, some papers that addressed specific topics were selected for more detailed qualitative analysis using N-Vivo software. Three categories, defined by the sample of network analysis, were explored in a cross-analysis of the 52 selected articles. This analysis allows a deeper understanding of the categories (1) CSR, (2) AI, and (3) Social Media (SM), and helps to identify the main stage of development and maturity use of AI in the corporate environment. The results point out some patterns and concerns common to different authors when addressing these specific topics, as highlighted in Table 1.

Table 1: Thematic analysis from the sampled documents

Research topic	Contributions	Suggested readings
Corporate Social Responsibility (CSR)	Authors address the importance of ethical aspects and the responsible use of AI, especially on data privacy. Approach on the relationship between AI and corporate social responsibility, suggesting that AI can be less socially irresponsible than humans. Call for humanistic HRM and recruitment which can increase reliability in the processes.	*Du and Chie, 2021* *Krkač, 2019* *Ligeiro et al., 2024* *McBride et al., 2022* *Rąb-Kettler and Lehnervp, 2019*
Artificial Intelligence (AI)	Using digital technologies and AI to reduce corporate risk, increase the efficiency and impact of their CSR activities. AI promotes corporate sustainable development. Provide status of Innovative companies. Lack of regulations in AI use.	*Atanasov et al., 2023* *Chen et al., 2024* *Grishunin et al., 2023* *Hemphill, 2020* *Londoño-Cardozo and Pérez de Paz, 2021* *Tseng and Lin, 2024*
Social Media (SM)	Main approach on social media responsibility, addressing privacy concerns with AI-driven social media technologies. Need to establish rules for responsible AI governance with deep democratic debate. Need to increase the information disclosure and transparency to increase trustworthiness.	*Buhmann and Fieseler, 2023* *Nave and Lane, 2023* *Walker and Milne, 2024* *Wu et al., 2024* *Xie et al., 2023*

The main results from this analysis evidenced that: (1) Ethical issues are a constant topic in all papers that deal with the three categories, although there are no conclusive results on this topic; an active and in-depth discussion by society is still needed to find a consensus bias. (2) There is a consensus that the use of AI tools increases productivity, reduces operational risks and negative impacts (social and environmental), and generates a certain trust from the user's perspective. However, there is still a gap in the regulation of their use and concerns about the boundaries of using these tools without dehumanizing operations. (3) Transparency and reliability are still unresolved topics, with many gaps to be addressed, especially in users' data protection. Paradoxically, in some survey results, especially with social media users, the feeling is of transparency in corporate operations. Finally, (4) there is also a consensus among the various authors on the need for governance policies on the use of AI tools in HRM, as it is undeniable that these tools have positive aspects, and managers may leverage them to improve human resources performance. However, it is also unanimous that close and careful attention to ethical and moral issues is needed, as well as promoting social well-being. These are also the principles of CSR.

This analysis points out some important issues that can be used to identify the stage at which companies are using AI and the concerns that are still in the spotlight, such as the ethical implications of AI decision-making, the adequacy of data governance practices and the integration of CSR into AI strategies, as well as the professional qualification for responsible action in this technological environment.

3.4 Maturity Model

Maturity is usually related to the quality, competency, or level of sophistication of objects or processes. Maturity Models have been developed across different fields to qualitatively assess the maturity of a specific domain or progress toward a particular objective. Each maturity level marks a distinct phase of development along a desired or typical evolutionary trajectory (Becker et al., 2009). These levels are characterized by a set of widely accepted, generalizable criteria, by a set of indicators. The lowest level signifies the initial stage of development, while the highest level represents full maturity and high performance. This structure makes maturity models useful for comparing and identifying areas for improvement (de Bruin et al., 2005). Consequently, in this study, maturity and performance are viewed as closely related, with higher maturity reflecting higher commitment to CSR.

Maturity Models in AI use in the context of CSR and HR lack of studies, as no papers were found related to this topic in the Scopus database. Meanwhile, there is a model of AI maturity developed by the Argos Multilingual

(2024), which indicates means to assess the path of digital transformation in companies and it is used as a basis for defining the maturity levels of the proposed model, with proper adequacy to the main theme of the study.

4. The AI Maturity Progression Framework Proposition

The proposed maturity model encompasses multifaceted dimensions addressed in the literature, including but not limited to algorithmic fairness, transparency, accountability, privacy protection, and employee empowerment. Moreover, this paper examines the practical implications of adopting the maturity model approach in real-world organizational settings. It highlights the significance of leadership commitment, organizational culture, and stakeholder engagement in driving CSR-oriented AI initiatives in HRM. Furthermore, it underscores the importance of collaboration between HR professionals, data scientists, ethicists, and other relevant stakeholders to ensure the ethical design, deployment, and governance of AI systems.

4.1 The Relevant Outcomes From the Analyzed Literature

The proposed maturity model, shown in Figure 7, encompasses five levels, from Beginner, ad-hoc implementations of AI with minimal CSR focus to advanced stages, Leader, where CSR principles are fully integrated into AI-driven HRM processes. At each maturity level, organizations can identify specific criteria related to policy implementation, stakeholder involvement, transparency, and the ethical use of AI. By using this maturity model, companies can systematically advance their CSR practices, ensuring that their use of AI in HRM is not only effective but also socially responsible and aligned with broader organizational values.

By identifying the intensity of AI use, technical knowledge, usability, and the extent of responsibility for the use, it is also possible to identify the stage at which companies are in terms of digital maturity. AI implementation and use practices are strongly linked to actions regarding moral and ethical aspects, as they affect both the internal environment of organizations and local communities and throughout the supply chain in relationships with suppliers in all stages of production process. In this sense, responsible and democratic use can bring benefits.

It is important to emphasize that, given the changes that digital transformation brings to companies, current business models also undergo changes and must be rethought, paving the way for innovation and organizational entrepreneurship. The higher the level of digital maturity of companies, the greater the transformation and innovation in business models, also bringing new challenges for CSR, towards a CDR.

Furthermore, the assessment of the maturity level requires a set of indicators and a process of continuous evaluation, with identification of areas for improvement. This is a proposal for future research, which can be considered as Digital Function Deployment (DFD) and can assist in the step-by-step identification and assessment of the company's level of maturity, as well as its evolution over time.

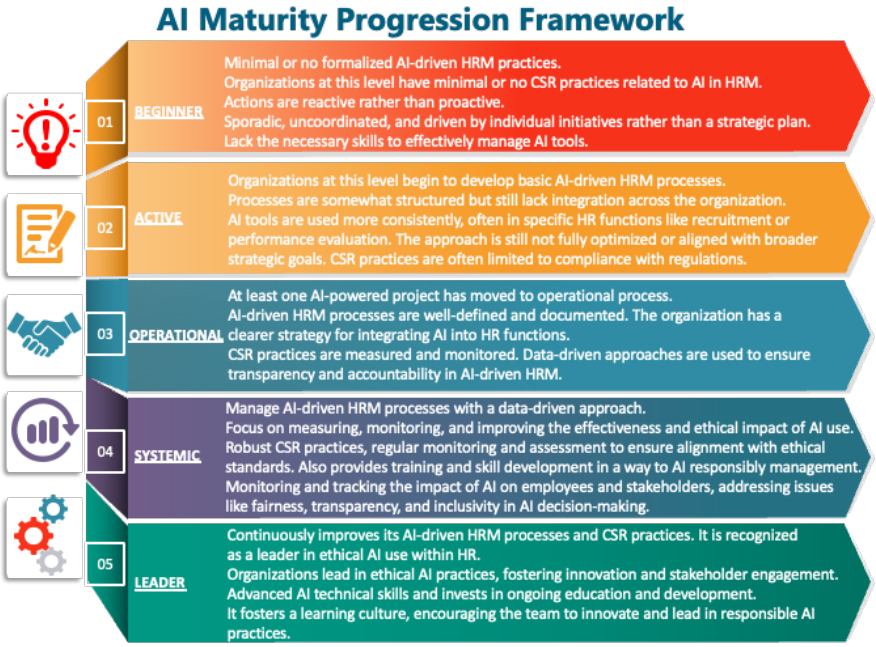


Figure 7: The AI Maturity Progression Framework (Source: Own elaboration)

5. Conclusion

This research addressed the gap between AI and CSR by exploring the literature on this edge and proposed a maturity framework to evaluate an organization's current state of AI integration in HRM practices. This framework also intends to guide the development of more responsible and ethical AI practices within the organizations' contexts. Among the findings, it was observed some interesting facts. The literature on the use of AI in HRM practices is still scarce, both from the perspective of companies and from the candidates' perception of the professional opportunities offered (Ligeiro et al., 2024). Many companies have made efforts toward the digital transition; however, the impacts of these changes are not yet known and raise questions about the humanization of recruitment and people management processes (Paul et al., 2024; Rab-Kettler and Lehnervp, 2019).

This research has multiple contributions. For organizations, this research sheds light on an emerging topic that is related to how AI is affecting recruitment through the lens of HRM. through the lens of HRM. For scholars, this research provides extensive suggestions for further research in the field of HRM and CSR, and how academia can help organizations foster CSR implementation in the light of AI. Additionally, this research highlights the ethical issues related to HRM and CSR due to the extensive spread in daily organizational activities. This paper is of interest to conference participants because it provides some highlights into the AI and CSR in the frame of HRM. It also provides to the attendees with actionable knowledge to foster AI and CSR within organizations, based on the importance of the skills needed.

Suggestions for future research can also be additionally addressed. Researchers can advance our knowledge about how to implement CSR practices in AI-driven HRM through different lenses like strategic orientation, policy development, training and employee development, stakeholder engagement, and continuous improvement at organizational levels. As a limitation, this research can present different results if simultaneous databases are summed to the results presented. This was due to the need for the input file to be from one source for CiteSpace calculations. Additionally, interpretations of qualitative analyses were made according to the authors' knowledge, which can vary from others' expertise. We hope this research encourages more scholars to explore the proposed gaps and advance the knowledge in these fields.

References

Akbarighatar, P. Pappas, I., Vassilakopoulou, P. (2023) "A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review" International Journal of Information Management Data Insights, Vol. 3, Nº 2, pp. 100193.

Aitken, M., Nh, M., Horsfall, D., Coopamootoo, K.P.L., van Moorsel, A., Elliott, K- (2021) "In pursuit of socially-minded data-intensive innovation in banking: A focus group study of public expectations of digital innovation in banking", Technology in Society, Vol. 66, pp. 101666.

Argos Multilingual. (2024) "The Argos AI Maturity Model: the 5 stages of the journey." Available at: https://info.argosmultilingual.com/ai-maturity-model. Access on July 8, 2024.

Atanasov, A., Chipriyanova, G., Krasteva-Hristova, R. (2023) "Integration of Digital Technologies in Corporate Social Responsibility (CSR) Activities: A Systematic Literature Review and Bibliometric Analysis", *Journal of Risk and Financial Management*, Vol. 16, Nº 8, pp. 373.

Becker, J., Knackstedt, R. and Pöppelbuß, J. (2009) "Developing maturity models for IT management", *Business and Information Systems Engineering*, Vol. 1 Nº 3, pp. 213-222.

Beliaeva, T., Ferasso, M., Kraus, S., Mahto, R. V. (2022) "Marketing and family firms: Theoretical roots, research trajectories, and themes", *Journal of Business Research*, Vol. 144, pp. 66-79.

Buhmann, A. and Fieseler, C. (2023) "Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence". *Business Ethics Quarterly.* Vol. 33, Nº 1, pp. 146-179.

Chang, Y. and Ke, J. (2024) "Socially Responsible Artificial Intelligence Empowered People Analytics: A Novel Framework Towards Sustainability, *Human Resource Development Review,* Vol. 23, Nº1, pp 88-120.

Chen, P., Chu, Z. Zhao, M. (2024) "The Road to corporate sustainability: The importance of artificial intelligence", *Technology in Society,* Vol. 76, pp. 102440.

Damoah, I.S., Ayakwah, A., Tingbani, I. (2021) "Artificial intelligence (AI)-enhanced medical drones in the healthcare supply chain (HSC) for sustainability development: A case study", *Journal of Cleaner Production*, 328, pp. 129598.

D'Cruz, P., Du, S., Noronha, E., Parboteeah, K.P., Trittin-Ulbrich, H., Whelan, G. (2022) "Technology, Megatrends and Work: Thoughts on the Future of Business Ethics", *Journal of Business Ethics*, Vol. 180, Nº 3, pp. 879-902.

de Bruin, T., Freeze, R., Kaulkarni, U. and Rosemann, M. (2005) "Understanding the main phases of developing a maturity assessment model", in Campbell, B., Underwood, J., Bunker, D. (Eds), Australasian Conference on Information Systems (ACIS), Nov. 30- Dec. 2, 2005, Sydney.

Du, S. and Xie, C. (2021) "Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities", *Journal of Business Research*, Vol. 129, pp. 961-974.

Ferasso, M., Beliaeva, T., Kraus, S., Clauss, T., Ribeiro-Soriano, D. (2020) “Circular economy business models: The state of research and avenues ahead”. *Business Strategy and the Environment*, Vol. 29, pp. 3006–3024. https://doi.org/10.1002/bse.2554

Ferasso, M., Tortato, U., Ikram, M. (2023) “Mapping the Circular Economy in the Small and Medium-sized Enterprises field: An exploratory network analysis”, *Cleaner and Responsible Consumption*, Vol. 11, pp. 100149.

Grishunin, S., Naumova, E., Burova, E. Suloeva, S. Nekrasova, T. (2022) “The Impact of Sustainability Disclosures on Value of Companies Following Digital Transformation Strategies”. *International Journal of Technology*, Vol. 13, pp. 1432-1441.

Hemphill, T.A. (2020) “The innovation governance dilemma: Alternatives to the precautionary principle”. *Technology in Society*, Vol. 63, pp. 101381.

Keenan, J., Kemp, D., Owen, J. (2019) “Corporate responsibility and the social risk of new mining technologies”, Vol. 26, Nº 4, pp. 752-760.

Krkač, K. (2029) “Corporate social irresponsibility: humans vs artificial intelligence”, *Social Responsibility journal*, Vol. 15, Nº 6, pp 786-802.

Lega et al., (2013)

Ligeiro, N., Dias, I. and Moreira, A. (2024) “Recruitment and Selection Process Using Artificial Intelligence: How Do Candidates React?”, *Administrative Sciences*, Vol. 14, pp. 155.

Londoño-Cardozo, J. and De Paz, M.P. (2021) “Corporate Digital Responsibility. Foundations and considerations for its development”, *Revista de Administração Mackenzie*, Vol. 22, Nº 6, pp. 1-31.

McBride, R. Dastan, A., Mehrabinia, P. (2022) “How AI affects the future relationship between corporate governance and financial markets: a note on impact capitalism”, *Managerial Finance*, Vol. 48, Nº 8, pp. 1240-1249.

Nave, E. and Lane, L. (2023) “Countering online hate speech: How does human rights due diligence
impact terms of service?”, *Computer Law & Security Review,* Vol. 21, pp 105884.

Rab-Kettler, K. and Lehnervp, B. (2019) “Recruitment in the Times of Machine Learning”, *Management Systems in Production Engineering*, Vol. 27, Nº 2, pp. 105-109.

Struensee, V.S. (2021) “Analyzing Dilemmas Posed by Artificial Intelligence and 4IR Technologies Requires Using All Available Models, Including the Existing International Human Rights Framework and Principles of AI Ethics”, Available at https://ssrn.com/abstract=3873097.

Tseng, C. and Lin, S. (2024) “Role of artificial intelligence in carbon cost reduction of firms”. *Journal of Cleaner Production*, Vol. 447, pp. 141413

Walker, k., and Milne, G. (2024) “AI-driven technology and privacy: the value of social media responsibility”, *Journal of Research in Interactive Marketing*. Vol. ahead-of-print DOI: 10.1108/JRIM-02-2024-0072.

Wu, l., Chen, Z., Tao, W. (2024) “Instilling warmth in artificial intelligence? Examining publics’ responses to AI-applied corporate ability and corporate social responsibility practices”, *Public Relations Review*, Vol. 50, pp. 102426.

Xie, I., Ma, W., Ma, Y., Ren, Z., Tong, Z., Wang, Y. (2023) “A counterfactual thinking perspective of moral licensing effect in machine-driven communication: An example of natural language processing chatbot developed based on WeChat API”, *Technological Forecasting and Social Change*, Vol. 197, pp. 122926.

Bridging the Gap: Practical Challenges and Strategic Imperatives in Adopting Gen AI

Andrea Di Vetta
School of Management, Faculty of Business and Law, Anglia Ruskin University, Cambridge, UK

andrea.divetta@gmail.com
andrea.di-vetta1@aru.ac.uk

Abstract: In today's rapidly evolving business landscape, Artificial Intelligence (AI) stands as the proverbial 'elephant in the room,' profoundly shaping diverse sectors and contexts. While debates rage among policymakers, practitioners, and politicians about regulating AI's widespread use, it's undeniable that AI represents a long-awaited digital technology poised to revolutionize organizational performance. Amidst the post-COVID era, the allure of AI has intensified, yet the critical question lingers: Can firms effectively harness AI and other technologies, such as Generative AI and Large Language Models (LLMs), to enhance their existing systems? Through a systematic literature review, a clear correlation between firms' implementation of AI, including the cutting-edge Generative AI, and their ability to adapt to changing market dynamics, drive operational excellence, and unlock new avenues for growth are explored. Moreover, key drivers of AI and digital adoption, such as the imperative for data-driven decision-making, the quest for customer-centricity, and the drive for sustainable business practices are explored. Our research not only highlights the transformative potential of AI and digital technologies but also provides actionable insights for business leaders navigating the complexities of technology adoption. By understanding the motivations, challenges, and strategic imperatives driving firms' technology choices, including the integration of Generative AI and LLMs, organizations can chart a path towards sustainable growth and competitive advantage in the digital age. This study underscores the revolutionary impact of Generative AI in digital transformation, offering a comprehensive understanding of their role in shaping the future of business.

Keywords:AI integration, Digital transformation, Organisational performance, AI-Driven innovation, AI technologies

1. Introduction

Artificial Intelligence (AI) and digital technologies have become integral components of modern business strategies. The post-COVID era has seen an intensified focus on digital transformation, with firms increasingly turning to AI to drive efficiency, innovation, and competitive advantage (McKinsey & Company, 2024). The COVID-19 pandemic accelerated the adoption of digital technologies as organizations sought to adapt to new market realities and operational constraints. This paper explores the adoption of AI, specifically Generative AI and Large Language Models (LLMs), by contemporary firms, analysing their impact on business processes and outcomes. The concept of AI has evolved over decades, transitioning from theoretical frameworks to practical applications in various industries. Early AI research focused on symbolic reasoning and expert systems, but recent advancements in machine learning and neural networks have propelled AI into mainstream business applications (Russell and Norvig, 2020). Today, AI technologies such as Generative AI and LLMs offer unprecedented capabilities in data processing, content creation, and predictive analytics (Jöhnk et al., 2021).

Generative AI, which includes technologies like Generative Adversarial Networks (GANs) and transformer-based models such as GPT-4, has revolutionized the way businesses approach content creation and data analysis. These models can generate human-like text, create realistic images, and even compose music, providing businesses with new tools for innovation and efficiency (Brown et al., 2020). The integration of these technologies into business operations promises to transform various aspects of organizational performance, from marketing and customer service to supply chain management and strategic planning (Zhang and Lu, 2021). The adoption of AI is not merely a technological upgrade but a strategic imperative that can redefine business models, enhance customer experiences, and create new value propositions (Dwivedi et al., 2021). For instance, AI-driven analytics enable more accurate forecasting and decision-making, which are crucial for maintaining competitiveness in rapidly changing markets (Jarrahi, 2018). Additionally, AI-powered automation can streamline operations, reduce costs, and improve efficiency, allowing firms to allocate resources more effectively (Davenport and Ronanki, 2018).

Despite the potential benefits, the path to successful AI adoption is fraught with challenges. Ethical considerations, such as bias in AI algorithms and concerns about data privacy, pose significant hurdles that firms must navigate (Goodman and Flaxman, 2017). Moreover, organizational resistance to change and the lack of a clear strategic vision can impede the integration of AI technologies (Jöhnk et al., 2021). The shortage of skilled personnel capable of developing, implementing, and managing AI systems further complicates the adoption process (Uren and Edwards, 2023). Through a systematic literature review, this research seeks to shed light on

how firms can effectively integrate these technologies to achieve sustainable growth and competitive advantage. By examining the motivations, challenges, and strategic imperatives behind AI adoption, this study aims to provide actionable insights for business leaders. This paper will review the latest advancements in AI, explore the diverse motivations driving its adoption, and critically assess the challenges and strategic imperatives that businesses face. By drawing on a wide range of secondary academic sources and industry reports, this study aims to contribute to the ongoing discourse on digital transformation and provide practical recommendations for leveraging AI to drive organizational success.

2. Literature Review

This literature review explores the adoption of Artificial Intelligence (AI) technologies, particularly Generative AI and Large Language Models (LLMs), within business contexts. The review focuses on studies published between 2010 and 2024 to capture the most recent developments and relevant research on AI's impact on business strategies and operations. This timeframe was selected due to the rapid evolution of AI technologies during this period, which saw a transition from theoretical frameworks to practical applications across various industries (Dwivedi et al., 2021; Russell & Norvig, 2020). The literature review also includes reports from consulting firms such as McKinsey & Company and Gartner, as well as business media outlets like Forbes, to illustrate current industry trends and the perceived benefits of AI technologies. However, it is important to acknowledge that these sources may have a vested interest in promoting AI solutions due to their roles in providing AI-related consulting services and their reliance on advertising revenue from technology firms. Consequently, while these reports offer valuable industry perspectives, their findings should be interpreted with caution as they may not always present an unbiased view (Smith, 2023; Johnson, 2024). To ensure a balanced analysis, this review also relies heavily on peer-reviewed academic studies that provide empirical evidence and theoretical frameworks for understanding AI adoption (Dwivedi et al., 2021; Jöhnk et al., 2021).

The literature highlights several key themes, including the role of AI in enhancing operational efficiency and driving innovation, the ethical and social implications of AI adoption, and the strategic alignment required for successful AI integration (Jöhnk et al., 2021; Snyder, 2019). Studies consistently show that businesses adopting AI technologies experience significant improvements in process automation and decision-making capabilities, while also grappling with challenges such as algorithmic bias and data privacy concerns (Goodman & Flaxman, 2017; Crawford, 2021). Despite the extensive secondary research available, there is a recognized need for more empirical research in this area. Future studies should focus on primary data collection methods, such as surveys and interviews, to provide deeper insights into the real-world experiences and strategies of businesses implementing AI. Such empirical research is crucial for advancing the understanding of AI's impact and guiding effective adoption strategies, making it a key future direction for this field of study. Overall, this literature review provides a comprehensive overview of the current state of AI adoption in business, emphasizing both the opportunities and challenges associated with this rapidly evolving field.

2.1 AI and Digital Transformation

AI technologies, including machine learning and natural language processing, have demonstrated significant potential in enhancing organizational performance. They enable automation, improve decision-making, and facilitate innovative solutions to complex problems (Dwivedi et al., 2021). The integration of AI in business processes has been linked to increased efficiency, cost savings, and improved customer experiences (Forbes, 2024). Research by Dwivedi et al. (2021) highlights the transformative impact of AI on business processes, noting that AI-driven automation can lead to substantial cost reductions and productivity gains. This assertion is supported by other studies, such as Brynjolfsson and McAfee (2017), who emphasize the role of AI in driving economic growth by enhancing productivity and creating new market opportunities. AI's role in digital transformation extends beyond operational efficiency. It includes the strategic alignment of technology with business goals to drive innovation and competitive advantage. According to Bessen (2019), firms that effectively integrate AI into their strategic planning processes are better positioned to leverage its full potential. AI technologies, particularly those that involve predictive analytics, enable businesses to anticipate market trends and customer preferences, thereby facilitating proactive decision-making and strategic agility (Jarrahi, 2018).

Generative AI and LLMs represent the latest advancements in AI technology. Generative AI can create new content, such as text, images, and music, based on existing data, while LLMs, like GPT-4, can understand and generate human-like text (Gartner, 2024). These technologies have been leveraged for tasks ranging from automated content creation to advanced data analysis, offering unprecedented opportunities for innovation and efficiency (Zhang and Lu, 2021). The evolution of these technologies has seen rapid advancements, with increasing applications in various industries, including marketing, healthcare, and finance (Jöhnk et al., 2021). In

healthcare, for example, LLMs are used to analyse patient data, predict disease outbreaks, and personalize treatment plans, thus improving patient outcomes and operational efficiency (Esteva et al., 2019). In finance, generative AI is employed to detect fraudulent transactions, develop personalized financial advice, and automate trading strategies (Boehme et al., 2020). Furthermore, a study by Jarrahi (2018) discusses how AI applications in business intelligence can enhance strategic decision-making capabilities.

2.2 Advantages and Applications of AI

Studies have shown that AI adoption leads to significant improvements in efficiency and decision-making. For example, AI-driven analytics enable more accurate forecasting and strategic planning (Jarrahi, 2018). Additionally, AI applications in customer service, such as chatbots, have enhanced customer satisfaction and engagement (Huang and Rust, 2018). Chatbots, powered by natural language processing, can handle a large volume of customer inquiries simultaneously, providing immediate responses and freeing up human agents for more complex tasks. The ability of AI to process large volumes of data swiftly and accurately has transformed business operations, leading to smarter and more agile enterprises (Davenport and Ronanki, 2018). AI-driven supply chain management systems, for instance, optimize logistics and inventory management by predicting demand and identifying potential disruptions before they occur (Dubey et al., 2020). This predictive capability reduces costs and enhances the resilience of supply chains, which is crucial in an increasingly uncertain global market. A recent study by Dubey et al. (2020) also underscores the role of AI in supply chain management, highlighting its potential to improve supply chain visibility and responsiveness. By integrating AI technologies into supply chain operations, companies can achieve real-time monitoring and predictive analytics, enabling them to respond swiftly to changes in demand or supply conditions. This agility is particularly important in industries where supply chain disruptions can have significant financial and operational impacts.

2.3 Critical Perspectives and Challenges

While the benefits of AI are well-documented, it is crucial to explore the challenges and potential downsides. One significant issue is the ethical implications of AI, including concerns about bias, privacy, and job displacement (Goodman and Flaxman, 2017). Bias in AI algorithms can perpetuate and even exacerbate existing inequalities, leading to unfair treatment of certain groups (Barocas, Hardt, and Narayanan, 2019). For example, facial recognition systems have been criticized for their higher error rates when identifying individuals from minority groups, raising concerns about their use in law enforcement and surveillance (Buolamwini and Gebru, 2018). Privacy concerns arise from the vast amounts of data required to train AI systems, raising questions about data security and user consent (Acemoglu and Restrepo, 2018). The collection and processing of personal data by AI systems can lead to unauthorized access and misuse of sensitive information, necessitating robust data protection measures and regulatory oversight. The General Data Protection Regulation (GDPR) in the European Union is an example of a legal framework aimed at protecting individuals' privacy in the context of data processing (Voigt and Von dem Bussche, 2017).

Moreover, the implementation of AI can be hindered by organizational resistance and the lack of a clear strategic vision. Firms may struggle with aligning AI initiatives with their overall business goals, leading to fragmented and ineffective efforts (Jöhnk et al., 2021). The shortage of skilled personnel to implement and manage AI technologies further complicates the adoption process (Uren and Edwards, 2023). The cost of implementing AI systems can be prohibitive for small and medium-sized enterprises (SMEs), limiting their ability to compete with larger organizations (Brynjolfsson and McAfee, 2014). Additionally, the dynamic nature of AI technology requires continuous learning and adaptation. Organizations must invest in ongoing training and development programs to ensure that their workforce is equipped with the necessary skills to leverage AI effectively (Dwivedi et al., 2021). Failure to do so can result in a knowledge gap that hinders the effective use of AI technologies.

2.4 An Ongoing Discussion

Some scholars argue that the hype around AI's potential benefits may overshadow its limitations. For instance, Brynjolfsson and McAfee (2014) caution against overly optimistic expectations, suggesting that the transformative impact of AI may be slower and less widespread than anticipated. They argue that while AI can automate specific tasks, it may not lead to significant overall productivity gains due to the complexity of integrating these technologies into existing workflows. Furthermore, there is a risk that reliance on AI could lead to complacency in human decision-making, with individuals placing undue trust in AI systems without critically evaluating their outputs (O'Neil, 2016). This could result in over-reliance on technology and a decrease in human analytical skills and judgment. Moreover, the overreliance on AI could stifle creativity and innovation by limiting the diversity of thought and problem-solving approaches (Crawford, 2021). AI systems, while powerful, are

inherently limited by the data they are trained on and the algorithms that drive them. Without human oversight and critical thinking, there is a risk of missing nuanced or context-specific insights that AI might not capture. Another critical perspective is the potential economic impact of AI on employment. While AI can create new job opportunities, it can also displace existing jobs, particularly in sectors reliant on routine and repetitive tasks (Acemoglu and Restrepo, 2018). This displacement can lead to economic inequality and social unrest if not managed appropriately. Policies aimed at workforce retraining and education are essential to mitigate the adverse effects of AI on employment (Autor, 2015).

In conclusion, while AI holds significant promise for transforming business processes and driving innovation, it is not without its challenges and limitations. A balanced approach that considers both the benefits and the ethical, social, and economic implications of AI is crucial for its successful adoption and integration into business strategies.

3. Methodology

3.1 Research Design and Strategy

This study employs a systematic literature review methodology to comprehensively examine the current state of knowledge regarding the adoption of Artificial Intelligence (AI), particularly Generative AI and Large Language Models (LLMs), in business contexts. The systematic review approach was chosen due to its structured nature, which facilitates the rigorous identification, evaluation, and synthesis of relevant studies, thereby providing a thorough understanding of existing research and highlighting key trends, challenges, and gaps in the literature (Snyder, 2019; Kitchenham & Charters, 2007). To identify relevant literature, a comprehensive search strategy was developed, encompassing multiple reputable academic databases such as Scopus, IEEE Xplore, and Google Scholar. These databases were selected based on their extensive coverage of multidisciplinary research, particularly in fields related to technology, management, and business. The search process involved the use of specific keywords and phrases, including "AI adoption in business," "Generative AI," "Large Language Models," "digital transformation," "AI-driven innovation," and "organizational performance." Boolean operators (AND, OR) were employed to refine and expand the search queries to capture a wide array of studies pertinent to the topic (Boell & Cecez-Kecmanovic, 2015).

A set of predefined inclusion and exclusion criteria was established to ensure the relevance and quality of the studies selected for review. Studies were included if they met the following criteria: (1) published in peer-reviewed journals between 2010 and 2024, (2) focused on the adoption or application of AI technologies in a business or organizational context, and (3) provided empirical or theoretical insights into the factors influencing AI adoption, the benefits and challenges associated with AI technologies, or strategic recommendations for AI integration. Studies were excluded if they were not peer-reviewed, such as conference papers, opinion pieces, editorials, or articles unrelated to the business applications of AI, to maintain a high standard of academic rigor and relevance (Tranfield, Denyer, & Smart, 2003).

3.2 Data Extraction and Analysis

The selected studies underwent a rigorous data extraction process, which involved categorizing the literature into key themes based on their focus and findings. The primary themes identified included AI-driven efficiency and operational excellence, ethical and social implications of AI adoption, strategic alignment and innovation, and challenges and barriers to AI integration. Each study was critically appraised for its methodological approach, sample size, data collection methods, and relevance to the research questions. A thematic analysis was conducted to synthesize the findings across different studies, which involved identifying patterns, relationships, and divergences among the research outcomes (Braun & Clarke, 2006). This approach enabled a comprehensive synthesis of existing knowledge, highlighting both consensus areas and ongoing debates within the field of AI adoption in businesses. To ensure the credibility and reliability of the findings, a quality assessment of the selected studies was conducted using a standardised appraisal tool. This assessment focused on evaluating the methodological rigor of each study, including the clarity of research design, the appropriateness of data collection and analysis methods, and the validity and reliability of the results presented. Studies that failed to meet the quality criteria were either excluded from the final synthesis or were considered with caution, with their limitations clearly noted in the discussion (Higgins et al., 2019).

The synthesis of findings involved a narrative approach to integrate the results from the included studies into a cohesive framework that addresses the research objectives of this paper. The narrative synthesis provided a detailed examination of how different businesses are adopting AI technologies, the strategies employed, the challenges encountered, and the impact of AI on organizational performance. Additionally, this synthesis

facilitated the identification of research gaps and areas for future investigation, particularly concerning the ethical implications and strategic management of AI technologies in diverse business environments (Popay et al., 2006).

3.3 Limitations of the Review

While this systematic review provides a comprehensive overview of the current literature on AI adoption in business contexts, it is not without limitations. Firstly, the review relies solely on secondary data sources, which limits the ability to capture real-time, firsthand insights from businesses currently adopting AI technologies. Future research should include primary data collection methods, such as quantitative surveys to gather data on the extent and impact of AI adoption across different sectors, and qualitative interviews with industry leaders and decision-makers to explore deeper insights into the challenges, strategies, and experiences of firms actively engaging in AI integration. Additionally, expanding future research to include a broader range of sources and studies published in other languages would help capture more diverse insights and experiences.

3.4 Relevance to Ongoing Work

This study is highly relevant to ongoing efforts in the field of AI and digital transformation. By providing a comprehensive overview of current knowledge and identifying critical gaps, this research supports business leaders, policymakers, and researchers in making informed decisions about AI adoption. The strategic recommendations offered in this study are particularly valuable for organizations seeking to leverage AI technologies for competitive advantage. The findings also contribute to the academic discourse by highlighting the multifaceted nature of AI adoption. This includes not only the technological and operational aspects but also the ethical, social, and economic implications. The study emphasizes the need for a balanced approach that considers both the potential benefits and the inherent challenges of AI integration. Furthermore, the identification of key themes and trends in the literature provides a foundation for future research. Scholars can build on these insights to explore specific areas in greater depth, such as the long-term impact of AI on employment or the effectiveness of different ethical frameworks in AI governance.

4. Discussion of Findings

The literature review provides several key insights into the adoption of Artificial Intelligence (AI) technologies, particularly Generative AI and Large Language Models (LLMs), within contemporary business contexts. This section discusses the main findings from the review, focusing on the benefits and challenges of AI adoption, strategic recommendations, and the need for empirical research to further understand AI's impact on businesses.

Despite the extensive research on AI and digital transformation, there remain significant gaps in understanding the practical challenges and strategic imperatives specific to the adoption of Generative AI and Large Language Models (LLMs) within contemporary firms. Current literature primarily focuses on the technical advancements and potential benefits of these technologies (Accenture, 2024; Gartner, 2024), yet there is a paucity of empirical studies examining the real-world application, especially within different organizational contexts and industries. Moreover, while several studies highlight the importance of leadership and strategic alignment in AI adoption (Forbes, 2024; McKinsey & Company, 2024), there is limited research that delves into the specific barriers and enablers experienced by firms in diverse sectors. This gap in the literature underscores the need for more nuanced insights into how organizations navigate the complexities of AI integration, including the roles of organizational culture, skills shortages, and strategic planning. To address this gap, this study will employ a mixed-methods approach, incorporating both quantitative surveys and qualitative interviews with industry leaders and decision-makers. By conducting semi-structured interviews, the research aims to gather in-depth insights into the experiences, challenges, and strategies of firms actively engaged in AI adoption. These interviews will complement the survey data, providing a richer understanding of the factors influencing successful AI implementation and offering practical recommendations for other organizations. Given the current gaps in the literature and the initial findings from our research, I now turn to the specific insights gathered from our empirical investigations. The following sections present the detailed findings from the study. These findings provide a comprehensive view of how AI technologies are being adopted and integrated into contemporary firms, highlighting both the benefits and the challenges encountered along the way**.**

4.1 Efficiency and Competitiveness

One of the primary motivations for AI adoption is the pursuit of efficiency and competitiveness. Firms implementing AI report significant improvements in operational efficiency and cost savings. Automation of

routine tasks through AI has reduced the need for extensive human intervention, leading to faster and more accurate processes (Yang, Blount, and Amrollahi, 2024). For example, AI-driven systems can handle repetitive tasks such as data entry, invoice processing, and customer service inquiries, freeing up human employees to focus on more strategic activities. According to a study by Dwivedi et al. (2021), firms that adopt AI technologies can achieve operational efficiencies that lead to competitive advantages. Furthermore, AI technologies enable predictive maintenance in manufacturing and other industries. By analyzing data from sensors and machines, AI can predict equipment failures before they occur, reducing downtime and maintenance costs (Lee et al., 2018). This proactive approach to maintenance not only improves efficiency but also extends the lifespan of equipment, contributing to long-term cost savings.

4.2 Innovation and Agility

AI technologies enable firms to innovate and adapt to changing market conditions. Generative AI, for example, allows for the rapid development of new products and services, enhancing a firm's ability to respond to customer needs and market trends (Accenture, 2024). For instance, AI-driven design tools can generate multiple product prototypes based on user preferences and feedback, significantly reducing the time and cost associated with traditional design processes (Holzinger, 2016). Additionally, AI-driven insights support agile decision-making, helping firms to pivot quickly in response to emerging opportunities or threats (McKinsey & Company, 2024). Research by Jöhnk et al. (2021) suggests that AI adoption facilitates organizational agility, enabling firms to swiftly adapt to market changes and innovate continuously. AI technologies provide real-time data analytics and forecasting capabilities, allowing businesses to make informed decisions rapidly. This agility is particularly valuable in highly competitive and dynamic markets where the ability to respond quickly to changes can be a significant competitive advantage.

4.3 Market Adaptation and Growth

The implementation of AI, particularly Generative AI and LLMs, has been correlated with improved market adaptation and growth. Firms leveraging these technologies can better anticipate market dynamics, optimize supply chains, and personalize customer interactions, leading to enhanced customer satisfaction and loyalty (Forbes, 2024). AI-powered recommendation systems, for example, analyze customer behavior and preferences to provide personalized product recommendations, thereby increasing sales and customer retention (Zhang et al., 2019). Dubey et al. (2020) highlight that AI-driven supply chain optimization can significantly enhance a firm's ability to respond to market demands, thus driving growth. AI technologies enable real-time tracking and analysis of supply chain data, helping companies to optimize inventory levels, reduce lead times, and improve overall supply chain efficiency. This capability is particularly important in industries such as retail and manufacturing, where supply chain disruptions can have significant financial and operational impacts.

4.4 Ethical and Social Implications

The adoption of AI also raises significant ethical and social considerations. Issues such as data privacy, algorithmic bias, and the potential for job displacement are critical concerns that must be addressed (Goodman and Flaxman, 2017). Bias in AI algorithms can perpetuate and even exacerbate existing inequalities, leading to unfair treatment of certain groups (Barocas, Hardt, and Narayanan, 2019). For instance, facial recognition systems have been shown to have higher error rates for minority groups, raising concerns about their use in law enforcement and surveillance (Buolamwini and Gebru, 2018).

Privacy concerns arise from the vast amounts of data required to train AI systems, raising questions about data security and user consent (Acemoglu and Restrepo, 2018). The collection and processing of personal data by AI systems can lead to unauthorized access and misuse of sensitive information, necessitating robust data protection measures and regulatory oversight. The General Data Protection Regulation (GDPR) in the European Union is an example of a legal framework aimed at protecting individuals' privacy in the context of data processing (Voigt and Von dem Bussche, 2017). The potential for job displacement is another critical concern. AI technologies can automate tasks that were previously performed by humans, leading to job losses in certain sectors. Acemoglu and Restrepo (2018) discuss the potential for AI to displace jobs, particularly in sectors reliant on routine tasks, highlighting the need for policies that support workforce retraining and adaptation. Without adequate measures to retrain displaced workers and prepare them for new roles, the economic and social impacts of AI-driven automation could be severe. Crawford (2021) argues that the ethical implications of AI require a multidisciplinary approach, involving stakeholders from various fields to develop comprehensive and inclusive solutions. This includes not only technical experts but also ethicists, policymakers, and representatives

from affected communities. Such an approach can help ensure that AI technologies are developed and deployed in ways that are fair, transparent, and accountable.

4.5 Strategic Recommendations

Based on the findings of the literature review, the following recommendations are proposed for businesses considering the adoption of Artificial Intelligence (AI) technologies, particularly Generative AI and Large Language Models (LLMs). These recommendations aim to guide organizations in effectively integrating AI into their operations and strategic planning to maximize the benefits while minimizing potential risks and challenges.

1. Develop a Clear Strategic Vision: Aligning AI initiatives with broader business objectives is crucial for cohesive and effective implementation. Firms must ensure that AI projects are not isolated efforts but integrated into the overall strategic vision. This involves setting clear goals, defining key performance indicators (KPIs), and regularly reviewing progress to align AI initiatives with business outcomes (Jöhnk et al., 2021). For example, a company might set a goal to improve customer satisfaction by 20% through the implementation of AI-driven customer service tools. By tracking KPIs such as customer response times and satisfaction ratings, the company can measure the impact of AI on its strategic objectives.
2. Invest in Skills and Training: Addressing the skills gap is essential for successful AI adoption. Firms should invest in training programs to develop internal expertise in AI technologies. Partnerships with academic institutions can facilitate access to cutting-edge research and emerging talent. Additionally, continuous learning and development programs can help employees stay abreast of the latest advancements in AI and apply them effectively within the organization (Dwivedi et al., 2021). For instance, offering courses on machine learning, data science, and AI ethics can equip employees with the knowledge and skills needed to leverage AI technologies effectively.
3. Implement Ethical Guidelines: Establishing ethical guidelines is critical to ensure responsible AI use. This includes developing frameworks to address issues such as bias, transparency, and data privacy. Firms should adopt practices that promote fairness, accountability, and transparency in AI decision-making processes. Implementing robust data governance policies can also help mitigate risks associated with data breaches and ensure compliance with regulatory requirements (Goodman and Flaxman, 2017).

5. Conclusion

AI and digital technologies, particularly Generative AI and LLMs, hold transformative potential for modern businesses. By understanding the motivations, challenges, and strategic imperatives behind AI adoption, firms can navigate the complexities of technology integration and achieve sustainable growth and competitive advantage. This study underscores the importance of a strategic, data-driven approach to AI adoption, offering valuable insights for business leaders. Future research should focus on longitudinal studies to observe the long-term impacts of AI adoption on business performance and market dynamics. Additionally, there is a need for further exploration of the ethical and social implications of AI, ensuring that its benefits are realized while mitigating potential risks. Policymakers and industry leaders must work collaboratively to establish standards and frameworks that promote responsible AI use and address the broader societal impacts of these technologies.

References

Accenture (2024). Reinvention in the age of generative AI. Available at: https://www.accenture.com/content/dam/accenture/final/accenture-com/document-2/Accenture-reinvention-in-the-age-of-generative-AI-executive-summary.pdf

Acemoglu, D., & Restrepo, P. (2018). Artificial Intelligence, Automation, and Work. *Journal of Economic Perspectives, 33*(2), 3-30.

Autor, D. H. (2015). Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives, 29*(3), 3-30.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press.

Bessen, J. (2019). AI and the Economy. *Journal of Economic Perspectives, 33*(2), 3-30.

Boell, S. K., & Cecez-Kecmanovic, D. (2015). On being 'systematic' in literature reviews. *Journal of Information Technology, 30*(2), 161-173.

Boehme, R., Claffy, K., Fontugne, R., & Fukuda, K. (2020). The CAIDA UCSD Anonymized Internet Traces - A 20-Year Retrospective. *Computer Communications Review, 50*(3), 5-21.

Braun, V., & Clarke, V. (2006)**.** Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Brooks, I., & Weatherston, J. (1997). *The Business Environment: Challenges and Changes.* London: Prentice Hall.

Brown, T. B., et al. (2020)**.** Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* WW Norton & Company.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81*, 1-15.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* New Haven: Yale University Press.

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review, 96*(1), 108-116.

Doherty, N., & Delener, N. (2001). Chaos Theory: Marketing and Management Implications. *Journal of Marketing Theory and Practice, 9*(4), 66-75.

Dubey, R., Gunasekaran, A., Childe, S. J., Blome, C., & Papadopoulos, T. (2020). Big data and predictive analytics and manufacturing performance: Integrating institutional theory, resource-based view, and big data culture. *British Journal of Management, 31*(2), 341-361.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. *International Journal of Information Management, 57*, 101994.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine, 25*, 24-29.

Forbes (2022). Paradox: Artificial Intelligence helps solve but suffers from skills shortages. Available at: https://www.forbes.com/sites/joemckendrick/2024/02/26/7-paradoxes-of-artificial-intelligence/

Forbes (2024). Advantages of AI. Available at: https://www.forbes.com/advisor/in/business/software/advantages-of-ai/#:~:text=Helps%20in%20Improving%20Processes%20andworkflows%20which%20are%20error%2Dfree

Forbes (2024). 11 barriers to effective AI adoption and how to overcome them. Available at: https://www.forbes.com/sites/bernardmarr/2024/05/10/11-barriers-to-effective-ai-adoption-and-how-to-overcome-them/

Gartner (2024). Generative AI is now the most frequently deployed AI solution in organizations. Available at: https://community.quantexa.com/discussion/4262/gartner-generative-ai-is-most-frequently-deployed-ai-solution#:~:text=Generative%20artificial%20intelligence%20(GenAI)%20isa%20new%20survey%20by%20Gartner

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine, 38*(3), 50-57.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane Handbook for Systematic Reviews of Interventions.* Wiley-Blackwell.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics, 3*(2), 119-131.

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research, 21*(2), 155-172.

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577-586.

Jöhnk, J., Weißert, M., & Wyrtki, K. (2021). Ready or not AI comes—An interview study of organizational AI readiness factors. *Business & Information Systems Engineering, 63*(1), 5-20.

Kitchenham, B., & Charters, S. (2007)**.** *Guidelines for performing systematic literature reviews in software engineering.* EBSE Technical Report.

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2018). Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. *Manufacturing Letters, 18*, 20-23.

McKinsey & Company (2024). In digital and AI transformations, start with the problem, not the technology.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown Publishing Group.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. *A Product from the ESRC Methods Programme.*

Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Harlow: Pearson.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333-339.

Tranfield, D., Denyer, D., & Smart, P. (2003)**.** Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207-222.

Uren, V., & Edwards, J. S. (2023). Technology readiness and the organizational journey towards AI adoption: An empirical study. *International Journal of Information Management, 68*, 102588.

Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide.* Springer.

Yang, J., Blount, Y., & Amrollahi, A. (2024). Artificial intelligence adoption in a professional service industry: A multiple case study. *Technological Forecasting and Social Change, 201*, 123251.

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration, 23*, 100224.

The Academic Anti-Procrastination Approach: Combining Peer Motivation and Personalized Artificial Intelligence Reminders

Xiaojiao Duan, Zhaoxia Yi, Yongjia Sun and Itamar Shabtai
Claremont Graduate University, Claremont, USA

Xiaojiao.duan@cgu.edu
Zhaoxia.yi@cgu.edu
Yongjia.suan@cgu.edu
Itamar.shabtai@cgu.edu

Abstract: Academic procrastination is a pervasive issue that significantly affects college students, leading to increased anxiety, stress, and reduced study efficiency and performance. Despite numerous studies exploring the causes and solutions to reduce procrastination, including the positive effects of peer motivation and technological interventions, the integration of artificial intelligence (AI) interventions through peer motivation and smart reminders remains underexplored. In this study, we conducted a systematic literature review on the causes of academic procrastination, the influence of social motivation, and technical interventions aimed at reducing procrastination. Our review revealed a significant research gap regarding the use of AI reminders and peer motivation to help students mitigate procrastination and enhance productivity. To address this gap, we propose an Academic Anti-Procrastination Approach that integrates peer motivation, social interaction, and AI-driven reminders. This approach utilizes the social networks of college students and incorporates AI tools to create a support system designed to reduce procrastination. We conducted an experiment to evaluate the effectiveness of this approach, using a mixed-methods methodology to analyze the results. Our findings suggest that the approach effectively reduces academic procrastination by harnessing the synergistic effects of peer motivation and AI-driven interventions. Quantitative results showed a p-value of 0.0017 in the experimental group, indicating a statistically significant decrease in procrastination scores after the intervention. Qualitative semi-structured interview results revealed that all participants found the personalized AI reminders helpful, with 87% stating that social motivation and interaction motivated them to complete tasks. Additionally, 80% of participants indicated that the concepts behind the approach would be useful in combating procrastination and expressed a willingness to use the approach and join a peer group to reduce their academic procrastination. This study offers practical contributions for combating academic procrastination in college environments. Students can utilize this method to create supportive peer groups and leverage personalized technological support, helping them overcome low efficiency and maintain focus on academic tasks.

Keywords: Procrastination, Peer motivation, Artificial intelligence, Reminders, Social Interaction, College students

1. Introduction

Procrastination is a common issue across the globe which affects most people regardless of occupation or social status. Avoiding procrastination is reported to be one of the top goals around the world. For instance, a meta-analysis by Professor Steel (2007) at the University of Calgary claims that 80 to 95 percent of college students procrastinate, particularly when it comes to doing their homework or coursework. He stated that procrastination can be defined and characterized as a self-regulatory failure leading to poor performance and reducing human well-being. Longitudinal studies have revealed that procrastination is positively associated with anxiety and negatively related to self-efficacy for self-regulated learning (Yerdelen et al., 2016).

Numerous studies have explored the internal and external causes of procrastination, including individual factors and influences from groups. Research on procrastination among college students reveals its negative relationship with self-esteem, self-efficacy, and motivation (Cuncolkar & D'Silva, 2020; Chen et al., 2016; Vij & Lomash, 2014). Social networks can trigger procrastination through persuasive design features (Alblwi et al., 2019), while group work with interdependence can decrease procrastination, especially for high trait procrastinators (Koppenborg & Klingsieck, 2022).

In the context of academic procrastination among college students, personal emotions and behaviors are mutually influenced by the encouragement and support from peers. College students' thoughts and behaviors are easily influenced by their peer groups. Peer attachment, motivation and support play crucial roles in mitigating academic procrastination (Jin et al., 2019).

Although substantial research has focused on identifying the causes of procrastination and proposing solutions, the potential of AI technology to enhance peer interaction and encouragement, thereby altering individual procrastination behavior, remains unexplored. This paper aims to design an approach to assist college students in overcoming academic procrastination by joining groups with common goals. By leveraging peer motivation

and personalized AI reminders, the approach seeks to reduce procrastination, provide timely encouragement throughout task completion, and ultimately improve productivity.

2. Systematic Literature Review

Academic procrastination is a prevalent issue among college students, significantly impacting their academic performance and overall well-being. This literature review aims to fill this gap by providing a detailed overview of research on academic procrastination among college students, focusing on its causes, the influence of peer motivation, and the potential of technical interventions to mitigate procrastination. This literature review seeks to address the following research questions:

- What are the primary causes of academic procrastination among college students?
- How does peer motivation influence academic procrastination?
- What technical interventions have been proposed and evaluated to reduce academic procrastination in college students?

2.1 Method and Search Strategy

The purpose of a systematic review is to answer specific questions based on an explicit, systematic, and replicable search strategy, with inclusion and exclusion criteria identifying studies to be included or excluded (Gough, Oliver & Thomas, 2017). Data are then coded and extracted from included studies to synthesize findings and shed light on their application in practice, as well as on gaps or contradictions. An exhaustive search was conducted using Google Scholar, Semantic Scholar, Web of Science, Zotero, and Elicit. We used terms such as procrastination/anti-procrastination, academic procrastination, group work, social motivation, peer influence, study productivity, college students, and other similar terms. The search technique was designed to locate papers related to the three research topics and questions. In total, 30 articles were used on the topic of academic procrastination among college students.

2.2 Literature Review Findings and Discussion

Causes of Academic Procrastination in College Students

Internal Causes: we investigated the internal causes of academic procrastination of college students. The findings show that procrastination is often linked to poor self-regulation, where students struggle to manage their time and prioritize tasks effectively (Steel, 2007). Low self-esteem and self-efficacy are significant predictors of procrastination, as students with low confidence in their abilities are more likely to delay tasks (Guo & Chen, 2024). Intrinsic motivation and self-determined extrinsic motivation are negatively correlated with procrastination, whereas amotivation and external pressure contribute to higher procrastination rates (Deci & Ryan, 1985).

External Causes**:** We found that the external factors leading to academic procrastination are primarily focused on the challenge and difficulty of the tasks, as well as the disruptive elements of the external environment. The external factors leading to academic procrastination are primarily focused on tasks perceived as aversive or overly challenging and are more likely to be procrastinated (Blunt & Pychyl, 2000). Distractions in the environment, such as social media and other online activities, significantly contribute to procrastination (Guo & Chen, 2024).

Peer Influence on Academic Procrastination

Peer conformity can have both positive and negative effects on academic procrastination. Peers who engage in procrastination can reinforce this behavior in others, while supportive peers can help reduce procrastination through encouragement and accountability (Mackinnon et al., 2016). Group work has been shown to reduce procrastination by promoting accountability and shared goals. However, it can also increase feelings of guilt and shame if group members do not meet expectations (Wolters, 2003). Social motivation, including encouragement from peers, is a crucial factor in reducing procrastination. Social networks provide support and motivation, helping students to stay on track with their academic tasks (Schmitz & Wiese, 2006).

Technical Interventions for Reducing Procrastination

Nudge Theory was originated by Richard Thaler and Cass Sunstein in the book Nudge Theory. Nudges, such as reminders and prompts, have been effective in reducing procrastination by keeping students focused on their tasks (Thaler & Sunstein, 2008). Tools that facilitate social interaction and collaboration, such as online forums and study groups, can help reduce procrastination by providing social support and accountability (Hrastinski,

2009). Techniques such as goal setting (SMART goals), calendar integration, reminders, self-monitoring, and task motivation (commitment, rewards, milestones) are effective in managing procrastination. These strategies help students organize their tasks, stay motivated, and complete assignments on time.

2.3 Conclusion of the Findings and Research gap

Academic procrastination is a multifaceted issue influenced by internal and external factors, social motivation, and technological interventions. Addressing procrastination requires a comprehensive approach that includes enhancing self-regulation and motivation, leveraging peer influence positively, and utilizing technological tools to support students in managing their tasks effectively. By understanding and addressing these factors, educators and researchers can develop strategies to reduce procrastination and improve academic performance. Due to the research gap identified during the literature review, we proposed an approach to reduce academic procrastination among college students. Existing studies have not addressed how to use AI technology to combine peer motivation and personalized reminders to alter individual procrastination behavior. This unique contribution is the focus of our research.

3. Design

3.1 Design Science Research Methodology

We designed and built our artifact using the Design Science Research methodology, following the guidelines of the Information Systems research framework presented by Hevner et al. (2004), which offers a problem-driven approach to assist IS designers. Based on the Design Science Research (DSR) framework, we outline the environment, IT artifacts, and knowledge involved in developing an Academic Anti-Procrastination Approach. The figure emphasizes the relevance and rigor in connecting the environment, IT artifacts, and knowledge. The development process is cyclical, requiring ongoing evaluation and adjustments to refine the approach and enhance its effectiveness in combating academic procrastination. (Figure 1)

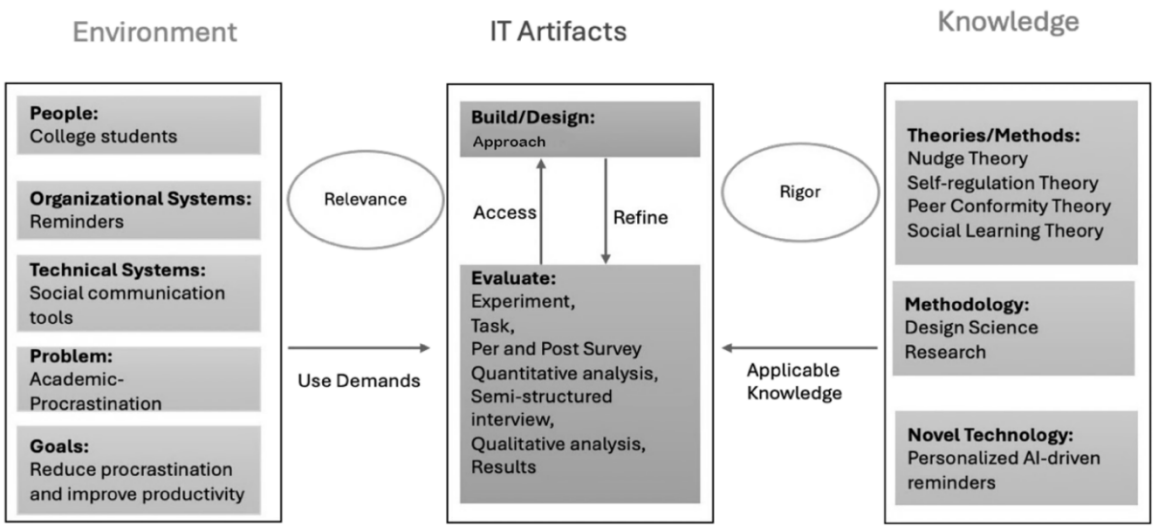


Figure 1: The Design Science Methodology applied in Academic Anti-Procrastination Approach

Environment: Based on the DSR model, our environment includes college students as the target population, digital reminders used during the learning process as the organizational systems, and social communication tools as the technical systems. The goal of this approach is to reduce academic procrastination and improve productivity.

IT Artifacts: The artifact is the "Academic Anti-Procrastination Approach," which integrates social motivation and AI-driven reminders. An experiment was designed to test the approach, including tasks, pre- and post-surveys for both experiment and control groups, and a mixed-methods approach involving quantitative analysis, semi-structured interviews, and qualitative analysis. The entire process is iterative, requiring continuous assessment and refinement.

Knowledge: To develop a robust approach for reducing academic procrastination, we integrated several foundational theories. Nudge Theory (Thaler & Sunstein, 2008) provided the basis for using subtle cues and reminders to steer students towards productive behaviors. Self-Regulation Theory (Schunk, 2012) emphasized

the importance of helping students manage their thoughts, emotions, and actions to achieve academic goals. Peer Conformity Theory (Laursen & Faur, 2022) highlighted the influence of peer behavior on individual actions, informing our approach to leverage peer motivation and social interaction. Lastly, Social Learning Theory (Brady, 2017) underpinned the framework's use of observational learning and modeling, encouraging students to learn from each other's successes and challenges. At the same time, we integrated cutting-edge AI technology into the process.

3.2 Academic Anti-Procrastination Conceptual Approach

To address academic procrastination among students, we developed a comprehensive approach that integrates several foundational theories with cutting-edge AI technology. This approach leverages personalized AI-driven reminders, peer motivation, and social interaction to enhance self-regulation and reduce procrastination. The integration of these constructs creates a multi-faceted approach that targets the key factors influencing procrastination. (see Figure2)

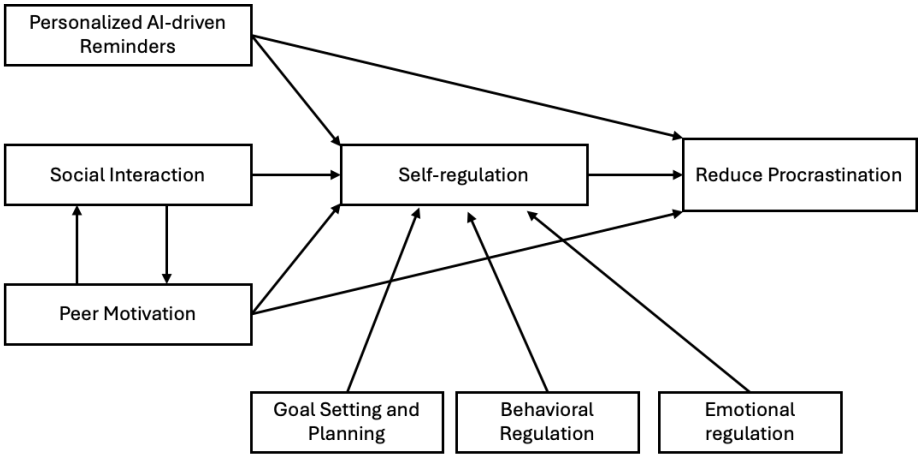


Figure 2: Academic Anti-Procrastination Approach

Variables in this Academic Anti-Procrastination Approach

Independent Variables*:* (a)Personalized AI-driven Reminders: tailored reminders generated by AI based on individual needs and preferences, aimed at keeping students on track. AI reminders provide timely cues and feedback to help students manage their tasks and time effectively. (b)Peer Motivation: encouragement provided by peers within the group which can foster a sense of community and mutual support, promoting motivation through social interactions. (c)Social Interaction: group discussions and collaborative activities designed to enhance engagement and motivation. Social communication can increase commitment to tasks through social support and interaction.

Dependent Variable: Reduction in Procrastination is the primary outcome of the approach, measured by the decrease in procrastination behaviors among students.

Mediator Variables: (a)Self-Regulation: the ability to manage one's thoughts, emotions, and behaviors in alignment with long-term goals. Acts as the central mechanism through which the independent variables influence the dependent variable. Enhanced self-regulation leads to reduced procrastination. (b)Goal Setting and Planning: establishing clear, achievable goals and creating structured plans to reach those goals. (c)Behavioral Regulation: controlling actions to avoid distractions and stay focused on tasks, using strategies such as time management and prioritization. (c) Emotional Regulation: managing stress and maintaining focus, such as mindfulness practices and relaxation exercises. Goal setting and planning, behavioral and emotional regulation directly contribute to self-regulation by ensuring effective self-management and focus on academic tasks.

4. Research Methodology

Our study employed a mixed-method approach within the Design Science Research (DSR) framework to develop and evaluate an Academic Anti-Procrastination Approach. We synthesized existing research on technological interventions and social motivation for reducing procrastination. Using DSR principles, we developed our approach integrating AI-driven reminders and peer motivation. We used Mixed-Method to evaluate this

approach, then synthesized quantitative and qualitative findings to comprehensively assess the approach's effectiveness.

By combining these methods, we aimed to provide a robust evaluation of our intervention's efficacy in reducing academic procrastination.

5. Evaluation and Results

5.1 Experimental Design and Implementation Process

To evaluate our approach, we conducted an experiment involving two groups: a treatment group and a control group. We utilized WhatsApp as the social communication tool for this experiment. A total of 34 participants, all university students from Claremont Graduate University, were randomly assigned to the two groups, with 17 participants in each group. The experiment comprised pre-surveys, post-surveys, completion of a specific academic task, and semi-structured interviews with the students in the treatment group. This approach enabled a comprehensive assessment of the approach's effectiveness in addressing academic procrastination.

The hypotheses for the experiment:

> *Null Hypothesis: Personalized AI-driven reminders combined with peer motivation and social interaction will not reduce procrastination.*
>
> *Hypothesis: Personalized AI-driven reminders combined with peer motivation and social interaction will reduce procrastination.*

Experiment Design: it was conducted over nine days (see Figure 4). On the first day, we sent the pre-survey to both groups and collected the responses by the end of the day. On the second day, we sent the "New Year's Study Planning Task" to both groups, giving them five days to complete it. During these days, based on each student's major, grade, and interests collected from the pre-survey of experiment group, we generated personalized AI reminders for each participant and encouraged the treatment participants to discuss and communicate about this task.

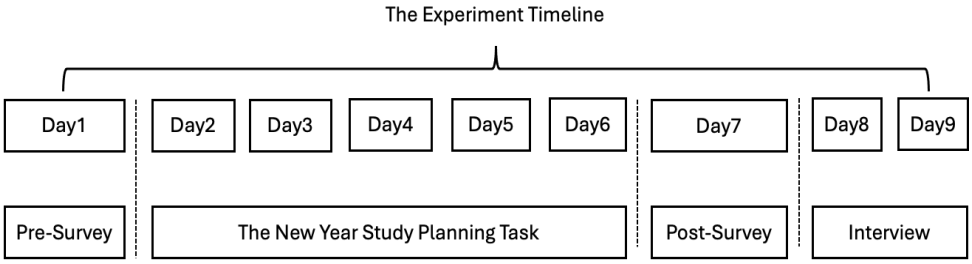


Figure 4: The Nine-Day Experiment Timeline

The pre and post survey: We used Qualtrics to distribute our surveys, which included both pre- and post-surveys for the control and experiment groups. The surveys, based on a 5-Point Likert Scale, consisted of four parts:

- Consent to Participate: it obtained participants' consent to take part in the study.
- General Procrastination Scale (GPS): The second part included the General Procrastination Scale (GPS), a reliable and validated measurement of procrastination containing 23 questions (Lodha et al., 2019).
- New Year Study Task and Attitudes: The third part consisted of four questions about the New Year study task and participants' attitudes toward peer motivation, social communication, and reminders during their study process.
- Personal Information: The final part concluded with personal questions about the participants' email address, major, grade, gender, interests, department, and school.

The task: named "New Year Study Plan," consisted of 10 open-ended questions about the past year's achievements and the new year's goals. Each student was given five days to complete this task. To measure the changes after the experiment, a post-survey, identical to the pre-survey, was administered. We will compare the General Procrastination Scale (GPS) scores before and after the task and analyze their attitudes toward peer motivation, social interaction, and AI reminders.

We began the experiment by adding the participants randomly into 2 groups, treatment and control. The participants in the treatment group could post in the chat. They also received reminders during the experiment. In contrast, the control group participants could not share anything and did not get any reminders during the experiment.

Personalized AI reminders: We sent educational, motivational, and recognitional AI reminders to the treatment group participants (see Table 1). The personalization of these reminders was achieved through a rule-based AI system that utilized individual student data collected from the pre-survey. The system followed these steps:

- Data Collection: The system gathered each student's major, grade level, and interests from the pre-survey responses.
- Reminder Template Selection: Based on the day of the experiment, the system selected appropriate reminder templates (educational, motivational, or recognitional).
- Content Customization: The AI then customized the content of each reminder by inserting relevant details such as the student's major, year of study, or specific interests into predefined placeholders within the templates.
- Timing Optimization: The system determined the optimal time to send reminders based on general student activity patterns and any time preferences indicated in the pre-survey.
- Adaptive Scheduling: The AI adjusted the reminder schedule based on task completion status, ceasing reminders once a student completed the task.

The treatment participants received personalized AI reminders each day. On the first day, each student received an educational reminder. From the second day to the fifth day, students received a mix of educational and motivational reminders randomly, at least one per day. Upon task completion, students received a recognitional reminder and no further reminders. The control group did not receive any reminders or encouragement posts, and there was no peer interaction within this group.

Table1: Personalized AI reminders Demo

Day	Educational Reminder	Motivational Reminder	Recognitional Reminder
Day 1	"Dear [Participant], next year is your [year] year at university, and your major is [major]. Start your study plan by reviewing your syllabus and your main tasks for the coming year. You will do your best!"		
Day 2	"Dear [Participant], set specific, measurable goals for each of your [major] subjects to keep your studies on track for your [year] year."	"Dear [Participant], remember, the future belongs to those who believe in the beauty of their dreams. Keep dreaming big!"	"Dear [Participant], you're doing amazing! Your effort today is a step towards success in [major]."
Day 3	"Dear [Participant], use active learning techniques like summarizing and questioning to enhance your understanding of [major] concepts."	"Dear [Participant], stay focused and stay strong. You're capable of achieving great things in your [year] year!"	"Dear [Participant], well done! You're halfway through your plan. Keep pushing forward in your [major] studies."
Day 4	"Dear [Participant], break down your tasks into manageable chunks to avoid feeling overwhelmed with your [major] coursework."	"Dear [Participant], believe you can and you're halfway there. Keep going!"	"Dear [Participant], your hard work is paying off. Great job staying dedicated to your [major] study plan."
Day 5	"Dear [Participant], review and adjust your study plan based on your progress to ensure you're on the right track in your [year] year of [major]."	"Dear [Participant], success is not final, failure is not fatal: It is the courage to continue that counts. You're almost there!"	"Dear [Participant], congratulations on completing your study plan! Celebrate your achievement and get ready for a successful year in [major]."

On the seventh day, we sent a link for the post-survey to both groups and collected the responses by the end of the day. Lastly, we conducted semi-structured interviews with the treatment group participants over the final two days. This structured approach allowed us to rigorously test the approach's impact on academic procrastination, providing both quantitative and qualitative data for a thorough evaluation.

5.2 Quantitative Evaluation

For the control group, 15 participants completed the pre-survey, 13 completed the task, and 13 completed the post-survey. In contrast, all 17 participants in the treatment group completed the pre-survey, the task, and the post-survey, resulting in a 100% completion rate. The completion rate for the control group was 76%. We performed t-tests using SPSS within the groups and between the groups, the statistics results as below:

Paired Samples T-Test for Experiment Group (N=16)

Paired Samples Statistic					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Pre-Test	65.12	17	7.98	1.94
	Post-Test	59.35	17	7.31	1.77

Paired Samples Test					
	Mean	Std. Dev.	t	df	P value
Pre-Post	5.76	5.23	4.53	16	<0.001

Paired Samples T-Test for Control Group (N=11, matched pairs only)

Paired Samples Statistic					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Pre-Test	61.14	14	8.59	2.30
	Post-Test	61.86	14	10.63	2.84

Paired Samples Test					
	Mean	Std. Dev.	t	df	P value
Pre-Post	-0.71	10.76	-0.25	13	0.808

Independent Samples T-Test (Comparing changes between groups)

Group Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Change	Experiment	5.76	17	5.23	1.27
	Control	-0.71	13	10.76	2.88

Independent Samples Test				
Change	t	df	P value	Mean Dif.
Equal Var. assumed	2.63	29	0.013	6.48

For the Experiment group, there's a significant decrease in procrastination scores (mean difference = 5.76, t (16) = 4.53, $p < 0.001$). For the Control group, there's no significant change in scores (mean difference = -0.71, t (13) = -0.25, $p = 0.808$). The pre and post scores are more strongly correlated in the experiment group than in the control group ($p = 0.013$), suggesting more consistent changes in the experiment group.

5.3 Qualitative Evaluation

After the treatment group completed their task and post-survey, we conducted semi-structured interviews with all 17 participants. Each participant answered seven questions. Using Atlas.ti for our qualitative analysis, we found that all participants considered the personalized AI reminders helpful. Additionally, 87% of participants felt that group communication and interaction motivated them to complete the task. Moreover, 80% appreciated the anti-procrastination approach and expressed interest in joining social groups to reduce procrastination in the future.

6. Conclusion

The Academic Anti-Procrastination Approach, which integrates peer motivation and personalized AI reminders, has demonstrated a significant impact on reducing procrastination among college students. The results from our mixed-methods approach provide robust evidence supporting the approach's efficacy. The quantitative analysis revealed a statistically significant decrease in procrastination scores within the experimental group. The paired samples t-test showed a mean difference of 5.76 ($t(16) = 4.53$, $p < 0.001$), indicating that the participants experienced a marked reduction in procrastination behaviors after the intervention. This significant decrease highlights the effectiveness of combining personalized AI reminders with peer motivation in tackling procrastination. The qualitative data from semi-structured interviews further validates the approach's success. All participants in the experimental group reported finding the personalized AI reminders helpful. 80% of the participants expressed a positive reception towards the anti-procrastination approach and showed interest in joining similar social groups to reduce procrastination in future academic endeavors. These findings underscore the practical applicability and user acceptance of the approach.

7. Limitations and Future Research

While our study shows promise in improving academic habits and potentially reducing procrastination, several limitations must be acknowledged: 1) Sample Size and Cultural Context: The experiment was conducted with a relatively small sample size of 34 participants from a single university in specify region/country. This limited sample size and specific cultural context may affect the generalizability of the findings. Cultural variations in understanding time management, peer support, and academic expectations could influence the effectiveness of our approach in different settings. 2) Task Design: The "New Year Study Plan" task, while useful for measuring procrastination reduction, may not fully represent the complexity and variety of typical academic assignments. The task's relevance may vary across cultures with different academic calendars or goal-setting practices.3) Time Constraints: The nine-day duration of the experiment limits our ability to fully evaluate the long-term effectiveness of the Academic Anti-Procrastination Approach. The differences in long-term planning and goal persistence could impact the approach's efficacy over extended periods. 4) Mediator Variables: The study did not explicitly test mediator variables such as self-regulation, goal setting, and emotional regulation, which play crucial roles in the approach and may be influenced by cultural factors.

To build on these findings and address limitations, future research should consider: a) Expanding the sample size and diversity, involving multiple universities across different cultural contexts to enhance generalizability. b) Designing longer-term experiments with more complex academic tasks that reflect various cultural academic environments. c) Exploring advanced AI functionalities that can adapt to cultural nuances in time management and academic expectations. d) Incorporating measures for mediator variables while considering their cultural relevance and interpretation.

By addressing these limitations and exploring research directions that consider cultural differences, this approach can be tailored to suit specific academic contexts, such as midterm or final exams, qualifying exams, and other high-level tasks. These scenarios often require enhanced peer motivation and AI technology to help students reduce procrastination and achieve their study goals. While our current findings are promising, further research is needed to develop a more robust and adaptable tool that can effectively support students in diverse cultural contexts. Ultimately, this approach has the potential to improve self-regulation skills, reduce procrastination, and contribute to better academic outcomes and overall well-being.

Acknowledgements

This study was supported by the CISAT department at Claremont Graduate University. The authors would like to express their sincere gratitude to Professor Itamar Shabtai for his valuable guidance, and to the researchers and student participants for their assistance in data collection. Their contributions were instrumental in the completion and success of this research.

References

Alblwi, A., Stefanidis, A., Phalp, K., & Ali, R. (2019). Procrastination on Social Networks: Types and Triggers. 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), 1-7.

Brady, C.M. (2017). Social Learning Theory.

Chen, B., Shi, Z., & Wang, Y. (2016). Do Peers Matter? Resistance to Peer Influence as a Mediator between Self-Esteem and Procrastination among Undergraduates. Frontiers in Psychology, 7.

Cuncolkar, S., & D'Silva, C.H. (2020). Relationship between Procrastination, Self-Esteem, Self-Efficacy and Motivation among College Students.

Gough, D., Oliver, S., & Thomas, J. (2017). An Introduction to Systematic Reviews (2nd ed.). Sage.

Guo, B., & Chen, J. (2024). The impact of social media addiction on academic procrastination among college students. Journal of Educational Psychology.

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research.

Hrastinski, S. (2009). A theory of online learning as online participation. Computers & Education.

Jin, H., Wang, W., & Lan, X. (2019). Peer Attachment and Academic Procrastination in Chinese College Students: A Moderated Mediation Model of Future Time Perspective and Grit. Frontiers in Psychology, 10.

Koppenborg, M., & Klingsieck, K.B. (2022). Social factors of procrastination: group work can reduce procrastination among students. Social Psychology of Education, 25, 249-274.

Laursen, B., & Faur, S. (2022). What does it mean to be susceptible to influence? A brief primer on peer conformity and developmental changes that affect it. International Journal of Behavioral Development, 46, 222-237.

Mackinnon, S. P., Sherry, S. B., & Pratt, M. W. (2016). The relationship between perfectionism, agency, and communion: A longitudinal mixed methods analysis. Personality and Social Psychology Bulletin, 39(11), 1520-1535.

Schmitz, B., & Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analysis of diary data. Contemporary Educational Psychology, 31(1), 64-96.

Schunk, D.H., & Zimmerman, B. (2012). Self-Regulation and Learning.

Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. Psychological Bulletin, 133(1), 65–94.

Thaler, R. H., & Sunstein, C. R. (2009). Nudge: Improving decisions about health, wealth, and happiness (Rev. and expanded ed). Penguin Books.

Vij, J., & Lomash, D.H. (2014). Role of Motivation in Academic Procrastination.

Wolters, C. A. (2003). Understanding procrastination from a self-regulated learning perspective. Journal of Educational Psychology, 95(1), 179-187.

Yerdelen, S., McCaffrey, A., & Klassen, R. M. (2016). Longitudinal examination of procrastination and anxiety, and their relation to self-efficacy for self-regulated learning: Latent Growth Curve Modeling. Kuram ve Uygulamada Eğitim Bilimleri/Educational Sciences: Theory & Practice, 16(1), 5–22.

How Good Is GPT's "Emojinal Intelligence"? Investigating Emoji Patterns in LLM-Generated Social Media Text

Michael Dunn and Kenneth Hopkinson
Air Force Institute of Technology, Wright-Patterson Air Force Base, USA

michael.dunn@afit.edu
kenneth.hopkinson@afit.edu

Abstract: Recent advancement in Large Language Models (LLMs) has opened the prospect of generating text for social media content that mimics human writing. The misuse of these tools presents urgent dilemmas, motivating the need to better understand the structure and patterns of LLM-generated content. Human communication on the Internet has developed relevant linguistic adaptations, including the use of emoji to augment traditional text. This study investigates the ability of one LLM, OpenAI's GPT-3.5, to replicate human emoji usage in social media contexts. Drawing upon a dataset of nearly three thousand US English human-written tweets, we employed GPT-3.5-Turbo to generate social-media-style content and analyzed the use of emoji in the resulting text. We compared the patterns of emoji usage between the LLM-generated and the human-written datasets, particularly frequency, types of emoji commonly used, and emoji sequences (n-grams). Our results revealed notable differences in all categories. While human-written tweets were more likely to use faces, hearts, and repetitive sequences of emoji, LLM-created content had a broader variety of emoji, with a preference for literal representations of the text's subject matter, producing diverse and unique emoji combinations.

Keywords: Large language models, Natural language processing, Computational linguistics, Social media, Emoji

1. Introduction

The recent advances in natural language generation by Large Language Models (LLMs) has enabled these models to produce text that closely resembles human-authored content. These LLMs, such as OpenAI's GPT-3.5 and GPT-4, are notable for both the quality and flexibility of their output, able to generate written compositions in a wide variety of styles and on nearly any topic as prompted by the user. This capability has and will continue to enable bad actors to use LLMs to speed up and potentially even automate detrimental content on social media, including spam, disinformation, and trolling. To combat such misuse, researchers and national security professionals need a better understanding of how LLM-generated content may appear in the online information environment.

Human language adapts remarkably well to changes in technology and communication media, and Internet-based communications platforms have been no different (McCulloch, 2019a). One language adaptation driven by the Internet, and especially social media, has been the invention and widespread adoption of emoji. These pictograms enable Internet users to add extra dimensions to their informal written communications ordinarily not possible with traditional text alone. Despite being quite a recent innovation, the use of emoji has already developed distinct and pervasive usage patterns. These are not rules to be followed but patterns that have emerged organically from humans experimenting, observing, sharing, and modeling.

The purpose of this work is to investigate the abilities of a current state-of-the-art LLM (GPT-3.5 Turbo) to mimic human emoji usage in social media content. We conduct an experiment to generate social-media-style content with the LLM, then compare the patterns of emoji use in the LLM-generated content to those of human social media users.

2. Background

Emoji, small pictograms that can be integrated into online text communication, were first created in Japan in 1999, and then introduced to the broader international community when they were made available on Apple's iPhone in 2011. Since text is limited in its ability to express tone, inflection, body language, and other embodied aspects of verbal communication, emoji have become an extremely popular method to encode additional meaning into informal digital communication. For a more thorough discussion of the history of emoji and their various linguistic properties, see McCulloch (2019).

Linguistics researchers have identified a number of patterns in the use of emoji in online communications. Gawne and McCulloch (2019) framed emoji usage as a form of digital gesture – a sort of linguistic property, but distinct from having grammar as a separate language. Their analysis demonstrated that the patterns of emoji usage are tightly analogous to those identified in the linguistics subfield of Gesture Studies. The gesture functions most relevant to our study here are illustrative, beat, and illocutionary.

Illustrative gestures are used to refer to physical objects or people in the real world, such as making the form of an item with the hands or indicating size with outstretched arms. Emoji used in this way depict graphically a thing or person in the real world – foods, sports gear, transportation vehicles, people in occupational clothing or engaging in an activity, etc. – and are generally used in a straightforward and literal manner. The majority of emoji options fall into this category, though it is not the most popular in actual usage. A property of this function is that there is no prescribed form. A sequence of birthday-related emoji will mean the same thing regardless of order or the specific order used (e.g. 🎉🎂🎈, 🎂🎈🥳, or 🎂🎉🎈).

Beat gestures are repetitive and rhythmic, used to emphasize important points of what the speaker is saying. A politician pounding a fist on the podium or a sports fan pumping an arm in the air while cheering are quintessential beat gestures. When watching a conversation between two people too far away to hear, it is very often possible to tell who is speaking, as they will move their hands and arms while speaking in a way that they do not while listening. Emoji usage patterns in online written communication are highly similar to beat gestures in verbal communication. Some are explicitly tied to physical gesture, such as repeated raised hands (🙌🙌) or clapping between/after words (WHAT 👏 IS 👏 GOING 👏 ON 👏) (Tatman, 2017). Most beat emoji, however, are repetition sequences, such as tears of joy (😂😂), heart eyes (😍😍😍), or loudly sobbing (😭😭😭😭). In fact, this is the most common type of emoji sequence. In a corpus of English-language users of the SwiftKey keyboard app, the top 9 emoji bigrams (two-character sequences) and the top 22 emoji trigrams and quadrigrams (three- and four-character sequences, respectively) were all pure repetition (McCulloch and Gawne, 2018).

Gestures that do not relate directly to the semantic meaning of the spoken or written words but rather "indicate the type of speech act" (Gawne and McCulloch, 2019) are described by some linguists as expressing the illocutionary force of the communication. That is, they give some indication of the speaker's intent, such as conviction, indifference, or sarcasm. Emoji used for this function are most often based on facial expressions or body language. For example, a smiling face (😊) might indicate pro-social intent, a smirk (😏) could indicate sarcasm, or a face throwing a kiss (😘) affection or flirtatiousness. As with gestures (McNeill, 2005, cited in Gawne and McCulloch, 2019), all of these functions that emoji can serve are non-exclusive, i.e. even a single usage can fall into more than one category. Emoji with illocutionary meaning can easily be repeated for emphasis as a beat gesture.

3. Methodology

The purpose of this experiment was to study emoji usage in LLM-generated social media text as compared to the patterns identified in the literature for human users. Since the emoji used are highly dependent on the intended meaning of the content, it was important that the subject matter of the LLM-generated text appropriately represented real human-written text on social media. To this end, all of the LLM-generated content was designed to mirror one-to-one the content of a representative dataset of human-written text. The dataset used is the UMass Global English on Twitter[1] Dataset, which contains over 10,000 randomly-sampled tweets, annotated with (among other things) English, non-English, and country (Blodgett et al., 2017; University of Massachusetts Amherst, 2017). Since the patterns of human emoji usage described in (McCulloch and Gawne, 2018; Medlock and McCulloch, 2016) are based on a corpus from users with their language set to US English, we focus our analysis initially on tweets annotated as English with country specified as US. This subset consists of 2,735 tweets.

Table 1: Examples of original tweets, topic, and a new tweet generated by GPT 3.5

Original tweet, from the UMass Global English on Twitter Dataset (Blodgett et al., 2017)	Topic, as summarized by GPT 3.5	LLM-generated tweet, from GPT 3.5 based on topic description
@[username] It's all about being real and true to yourself. The opposite of real is fake.	Authenticity and self-acceptance are key.	Just remember, being true to yourself and accepting who you are is the ultimate key to happiness and success. #authenticity #selfacceptance 💛🌟

[1] In July 2023, Twitter changed the name of its platform to X. Since both the dataset and most of the literature come from before this name change, for consistency this paper refers to said platform as Twitter.

Everything I don't want to happen is happening. Smh I hate my life right now .!😔	Negative feelings about current life circumstances.	Feeling super overwhelmed and discouraged by my current life circumstances right now 😔 Just trying to stay positive and remember that tough times don't last forever 💪 #perspective #staystrong
I just wanna sleep all day tomorrow!	Desire for extended sleep and relaxation.	Current mood: just want to crawl back into bed and sleep for like, three days straight 😴☁️ #needmorerest #relaxationgoals

Each of the tweets in this US English subset is then summarized by individually feeding through GPT 3.5 (specifically GPT-3.5-turbo-0125), using OpenAI's API (OpenAI, 2023), with a prompt to "describe the overall topic of this tweet in eight words or fewer."[2] The LLM-created content was then generated by prompting GPT 3.5 to write a tweet about each of those topics in turn (see Table 1 for examples). Only one artificial tweet was generated from the topic of each authentic tweet. Our dataset for analysis then is a set of 2,735 LLM-generated artificial English-language tweets. Using AntConc (version 4.2.4) corpus linguistics toolkit and regular expressions, we extracted list of the most common emoji n-grams for n=1, 2, 3, and 4. In addition to comparing LLM-generated tweets against the human patterns observed in the literature, we also look at emoji usage in the original dataset using the same extraction of common emoji n-grams.

Many emoji are encoded as multiple symbols that are then combined and rendered as one icon. This includes appending skin-tone markers for emoji depicting body parts and gender markers for depictions of persons (The Unicode Consortium, n.d.). Additionally, a special non-printable character called a Zero Width Joiner (ZWJ) can be used to combine emoji in different ways ("Zero Width Joiner Emoji," n.d.): some combinations create a person with a depicted profession, optionally with gender and skin tone markers; some create family groupings. For this analysis, all skin-tone and gender modified emoji are counted together with their unmodified base emoji. ZWJ groupings are treated as n-gram sequences of their component emoji, as ZWJ emoji groupings are relatively new and not uniformly supported.

4. Results

Across the 2,735 tweets generated by the LLM, there were 4,434 displayable Unicode characters associated with emoji[3] – an average of 1.6 emoji characters per generated tweet. While there are some overlaps between the list of top emoji in the LLM-generated tweets and the human-written tweets, they are largely very different distributions. Not only does the LLM's dataset contain vastly more emoji for the same number of tweets, but most of the top-20 common emoji (Table 2) are greatly overrepresented compared to human-written content. For instance, the most common emoji in the LLM set is raising hands (🙌), appearing in fully 6% of tweets; by contrast, this emoji appears in less than 0.4% of tweets in the human set. The flexed muscle arm (💪) is even more skewed. It was in only 6 tweets in the human set, but 133 LLM-generated tweets, more than 22 times as frequent! Eleven of the top 20 most common emoji in the human set of tweets are faces, including the top three (see Table 4); only four of the top 20 in the LLM set are faces, the highest ranked being only the 10th most frequent (😂). The results from the human-written tweets in this dataset are fairly consistent with emoji usage on Twitter overall. Of the top 20 most common emoji in the original human dataset 10 are also in the top 20 most popular emoji used globally on Twitter as of 2023, according to Emojitracker (Emojipedia, 2023).

Table 2: Most common emoji in LLM-generated tweets

Rank	Emoji	No.of tweets	Total no.of emoji
1	🙌	165	165
2	💕	138	138
3	💪	133	133
4	🎉	128	129

[2] We found in the methodology design process that if asked to describe the topic of a tweet without any restrictions, GPT 3.5 often would produce a summary as long or longer than the original tweet itself, and sometimes even repeat whole chunks of the input verbatim.

[3] This total includes 347 skin-tone and gender modifiers, and 64 "Regional Indicator Symbol Letters" that are used on compatible platforms in 2-character groupings to encode a single national or international flag (e.g. **U S** would be displayed as 🇺🇸).

Rank	Emoji	No.of tweets	Total no.of emoji
5	🔥	126	147
6	☀️	100	103
7	🎶	94	94
8	💼	93	99
8	✨	93	94
10	😂	77	81
11	🙄	72	74
12	😍	61	61
13	🤩	56	56
14	❤️	54	54
15	🚗	51	52
16	☀	42	42
16	🙏	42	42
18	🚨	41	50
19	💔	40	40
20	👩	39	39

Notes: N=2,735; ranked by the total number of tweets each emoji appears in; total no. of emoji includes some sequences of repeated emoji (see Table 3)

Table 3: Most common emoji n-gram sequences in LLM-generated tweets

Bigrams		Trigrams		Quadrigrams	
Emoji	No.	Emoji	No.	Emoji	No.
🤰💉	12	🔥🔥🔥	9	👱👩👧👱*	6
🔥🔥	11	👩👧👱*	7	🤰💼👩‡*	3
🎉🎂	11	👱👩👧*	6		
🍔🍟	10	🤰💉💼	4		
🔥🎶	8	🤰💼👩	4		
🤰💼	7	☀️👩💼*	3		
👧👱*	7	🍔🍟🍕	2		
👩👧*	7	🤰💉✨	2		
💼👩*	7	💕👫👫	2		
🚓🕶	7	🚓☂🕶	2		
☀️👩	6				
👱👩*	6				
💕✨	6				
😍👐	6				
🚗⚪	6				

Notes: Ranked by total number of tweets each sequence appears in; only n-grams appearing in at least 2 tweets are shown; * indicates ZWJ groupings.

The differences in multi-character sequences of emoji between LLM-generated text and human-composed is even more stark. Linguistics researchers studying patterns of emoji use online (by humans) have found that when multiple emoji are used in a sequence, it is overwhelmingly most common in repeated patterns (McCulloch and Gawne, 2018; Medlock and McCulloch, 2016). McCulloch and Gawne (2018) found that of the top 200 most common n-grams (n=2, 3, and 4) in their SwiftKey US English corpus, roughly half were strict repetition of identical emoji characters. At the top this trend is even stronger: nine of the top ten emoji bigrams, and all of the top ten trigrams and quadrigrams were strict repetition. Even among non-identical sequences, most show some form of repetition, such as different types of hearts (e.g. 💕 💖), or different faces expressing the same sentiment (e.g. 😍 😘). This is true of the UMass corpus (Blodgett et al., 2017) as well, with 15 of the top 25 most common emoji bigrams being straight repetition, and all but one of the emoji trigrams and quadrigrams appearing more than once being pure repetition. The LLM-generated tweets show a very different pattern of emoji use. Only one of the top 50 most common emoji bigrams and trigrams are identical repetitions. In fact, in the entire dataset of 2,735 LLM-generated tweets, only three repetition sequences appear in more than one tweet: 🔥 🔥 🔥 nine times, 🔥 🔥 twice, and 😂😂 twice. There are zero sequences of four repeated emoji that appear more than once in the entire LLM-generated corpus, and only one sequence of four or more repeated emoji at all, appearing just once (⭐ ⭐ ⭐ ⭐ ⭐, describing a movie). What we see instead are mostly small heterogeneous groupings, thematically matching the text, but not matching the repetition patterns prevalent in real human usage. The LLM had a strong tendency to use pairs of different emoji to reflect the topic literally, as with illustrative gestures. Use of emoji as beat or illocutionary features, both of which are much more common than illustrative uses among human writers, was noticeably lacking among the LLM-generated text.

Table 4: Most common emoji the original set of human-composed tweets

Rank	Emoji	No. of tweets	Total no. of emoji
1	😂	63	115
2	😍	35	49
3	😭	28	49
3	👌	28	33
5	❤	27	37
6	😘	20	24
7	😔	19	21
8	😩	17	22
8	💕	17	19
8	😊	17	19
11	😌	13	13
12	🤦	11	13
13	😳	10	13
13	😎	10	12
13	🙌	10	11
16	💙	9	15
16	💯	9	13
16	😋	9	11
19	🔥	8	12
19	🎉	8	9

Notes: N=2,735; ranked by the total number of tweets each emoji appears in; total no. of emoji includes sequences of repeated emoji (see Table 5).

Table 5: Most common emoji n-gram sequences in the original set of human-composed tweets

Bigrams		**Trigrams**		**Quadrigrams**	
Emoji	No.	Emoji	No.	Emoji	No.
😂 😂	21	😂 😂 😂	13	😂 😂 😂 😂	7
😭 😭	8	😭 😭 😭	7	😭 😭 😭 😭	2
😍 😍	7	😍 😍 😍	4	❤❤❤❤	2
❤ ❤	4	❤ ❤ ❤	3	♥♥♥♥	2
😂 😭	3	😱 😱 😱	2		
😱 😱	2	♥ ♥ ♥	2		
♥ ♥	2	🏈 🏈 🏈	2		
👌 👌	2	💯 💯 💯	2		
🏈 🏈	2	🔥 🔥 🔥	2		
💯 💯	2	😂 😭 😭	2		
🔥 🔥	2	😘 😘 😘	2		
😘 😘	2	😷 😷 😷	2		
😩 😩	2				
😷 😷	2				
🎃 👻	2				
😭 😂	2				
💰 💰	2				
😂 😍	2				
😊 😘	2				
😍 ❤	2				
😍 😘	2				
😩 😍	2				
😩 😭	2				
😭 ❤	2				
🙋 🙋	2				

Notes: Ranked by total number of tweets each sequence appears in; only n-grams appearing in at least 2 tweets are shown.

5. Conclusion

This experiment quantitatively analyzed how emoji are used in LLM-generated social media style content and compared that to the typical use by human social media users. As we demonstrated in section 4, tweets written by this LLM, OpenAI's GPT-3.5-Turbo, display markedly different patterns of emoji usage than humans. The most common emoji usage by humans consists largely of faces, hearts, and hands (Emojipedia, 2023), primarily to convey tone and intent as a substitute for body language. Repetition is typical, and by far the most common types of emoji sequences found in the wild are straight repetition. LLM-produced content, on the other hand, includes a much wider variety of emoji, with fewer faces and more literal representations of the text's subject matter. GPT 3.5 seems to favor a few emoji in particular, notably the raised hands (🙌) and flexed bicep (💪). Some of this overuse is likely due to the LLM's deliberate tuning by its developers. For instance, many uses of the flexed bicep emoji are in response to prompted topics expressing negativity, to which GPT 3.5 added an unprompted positive outlook (see Table 1 for one example).

In stark contrast to humans' proclivity for repetition using only a small subset of the most popular emoji, the LLM's emoji groupings were diverse and varied, resulting in many unique combinations rather than the same predictable few. This can result in the LLM-generated "tweets" looking a lot more like generic corporate brand accounts than average users (one example from the LLM content: "Don't forget to tune in to the big game tonight at 6:30 PM! Who's ready for some intense action on the field? 🏈 ⚽ #sports #excited #gametime" [A football and a soccer ball? Which 'big game' are we watching exactly?]).

These results have implications for the open question of detecting LLM-generated content on social media. While emoji use alone cannot flag an online post as artificial, patterns of emoji use over time could be used as one element contributing to detection of accounts posting large amounts of LLM-generated content.

References

Blodgett, S.L., Wei, J.T.Z. and O'Connor, B. (2017) 'A Dataset and Classifier for Recognizing Social Media English', *3rd Workshop on Noisy User-Generated Text, W-NUT 2017 - Proceedings of the Workshop*, pp. 56–61. doi:10.18653/V1/W17-4408.

Emojipedia (2023) *Emojitracker*. Available at: https://emojitracker.com/ (Accessed: 13 March 2024).

Gawne, L. and McCulloch, G. (2019) 'Emoji as Digital Gestures', *Language@Internet*, 17, p. article 2. Available at: https://www.languageatinternet.org/articles/2019/gawne/index_html (Accessed: 11 February 2024).

McCulloch, G. (2019a) *Because Internet: Understanding the New Rules of Language*. New York: Riverhead Books.

McCulloch, G. (2019b) 'Emoji and Other Internet Gestures', in *Because Internet: Understanding the New Rules of Language*. New York: Riverhead, pp. 155–196.

McCulloch, G. and Gawne, L. (2018) 'Emoji Grammar as Beat Gestures', in *Proceedings of the 1st International Workshop on Emoji Understanding and Applications in Social Media (Emoji2018)*. Available at: http://ceur-ws.org.

McNeill, D. (2005) *Gesture and Thought*. Chicago: University of Chicago Press.

Medlock, B. and McCulloch, G. (no date) *The Linguistic Secrets Found in Billions of Emoji - SXSW 2016 presentation*. Available at: https://www.slideshare.net/SwiftKey/the-linguistic-secrets-found-in-billions-of-emoji-sxsw-2016-presentation-59956212 (Accessed: 11 February 2024).

OpenAI (2023) *Text generation models*, *OpenAI developer platform documentation*. Available at: https://platform.openai.com/docs/guides/text-generation (Accessed: 6 December 2023).

Tatman, R. (2017) *Where 👏 do 👏 the 👏 claps 👏 go 👏 when 👏 you 👏 write 👏 like 👏 this 👏?*, *Making Noise & Hearing Things*. Available at: https://makingnoiseandhearingthings.com/2017/07/13/where-👏-do-👏-the-👏-claps-👏-go-👏-when-👏-you-👏-write-👏-like-👏-this-👏/ (Accessed: 3 March 2024).

The Unicode Consortium (no date) *Full Emoji Modifier Sequences, v15.1*.

University of Massachusetts Amherst (no date) *Twitter Language Identification: Data and Software*, *SLANG Lab at UMass*. Available at: https://slanglab.cs.umass.edu/TwitterLangID/ (Accessed: 4 March 2024).

Zero Width Joiner Emoji (no date). Available at: https://emojipedia.org/zero-width-joiner (Accessed: 4 March 2024).

Online Managerial Tools as Research Tools to Apply Artificial Management: Results of Research

Olaf Flak
Management Institute, Jan Kochanowski University of Kielce, Poland

olaf.flak@ujk.edu.pl

Abstract: Artificial intelligence (AI) can augment human intelligence in teamwork, however, it is still not clear how to implement artificial management. This rapid development of computer science gives opportunities to replace team managers with robots. However, it is still not possible to employ a robot in a managerial position. Therefore, the aim of this paper is to present a theoretical foundation for such an information system which could be widely used by human managers in their day-to-day work. At the same time, it could collect data on managerial work in order to implement artificial management. The research problem concerns the theoretical assumptions needed to design this solution. Two research questions arise from this research problem: (RQ1) What types of research methods should be used to collect data on managerial work? (RQ2) How to implement research methods into the managerial tools to make the managerial work automated? The research methods used in the paper are: literature studies, technical documentation of online managerial tools created by the author, and a long-term observation of human-managed teams. In Section 2 we present the theoretical foundation of research in management science which is the answer to the first research question (RQ1). Section 3 contains the theoretical assumption of the information system design which is the answer to the second research question (RQ2). In Section 4 we present examples of research conducted by the author as the practical use of answers to both research questions (RQ1 and RQ2). Section 5 includes conclusions on the use of the information system in artificial management. The main contribution of this paper is as follows: Firstly, to answer the first research question (RQ1) about the types of research methods which should be used to collect data on managerial work in order to make them automated. The answer is the mixed method as the most appropriate method to study what a manager really does. Secondly, the paper contains the answer to the second research question (RQ2) on ways of implementing research methods into the online managerial tools aimed at artificial management.

Keywords: Artificial management, Artificial intelligence, Information system, Research methods, Managerial work, Team management

1. Introduction

For the last 20 years there has been a rapid development of information technology, robotics and the replacement of people's work with machines or algorithms. Managers commonly work with electronic tools that facilitate their work whilst registering their work (Ewenstein, Hancock and Komm, 2016), e.g. in logistics processes (Dash, McMurtrey and Rebman, 2019), IT sector services (Keller, 2017) as well as team management (Davenport and Kirby, 2015). This rapid development of computer sciences gives opportunities to replace team managers with robots. However, it is still not possible to employ a robot in a managerial position.

Artificial intelligence (AI) can augment human intelligence in teamwork, however, it is still not clear how to implement artificial management (Flak and Pyszka, 2022). Research on AI shows that it could enhance human teams in the area of team coordination, knowledge sharing and learning, supporting decision making, as well as evaluation and team performance (Khakurel and Blomqvist, 2022). Additionally, the use of transformer-based language models in artificial intelligence (AI) has increased adoption in various industries and led to significant productivity advancements in business operations (Bouschery, Blazevic and Piller, 2023).

The idea of replacing human managers with robots has been shaped by the idea of an intelligent knowledge system in an organization (Davern, 2007). Nowadays, there are many information systems for knowledge management in organizations focused on sale, distribution or production (Yanchinda, et al., 2011). However, there is still very little research in the field of team management. Despite the fact that it is common to track activities of Internet users or some online applications (i.e. Google apps) and a wide theoretical background of knowledge systems in organizations and a large number of computer systems in companies (SAP, ERP), there is hardly any research based on online management tools in order to obtain data about managers' behaviours. It seems that there is a lack of a complex system of management tools for team management and collecting data for using pattern recognition of human managers' activities.

Therefore, the aim of this paper is to present a theoretical foundation of such an information system which, on the one hand, could be widely used by human managers in their day-to-day work, and on the other hand, it could collect data on managerial work in order to implement artificial management.

The research problem concerns the theoretical assumptions needed to design such an information system. Two research questions arise from this research problem:

RQ1: What types of research methods should be used to collect data on managerial work?

RQ2: How to implement research methods into the online managerial tools to make the managerial work automated?

The research methods used in the paper are literature studies, technical documentation of online managerial tools created by the author, and a long-term observation of human-managed teams.

In Section 2 we present a theoretical foundation of research in management science. Section 3 contains theoretical assumptions of the information system design which is the answer to the first research question (RQ1). In Section 4 we present examples of research results conducted by the author as the answer to the second research question (RQ2). Section 5 includes conclusions on the use of the information system in artificial management.

2. Theoretical Foundation of Research in Management Science

Saunders and Bezzina (2015) write that in organization theory (it seems also more broadly: in management science – author's note) two main epistemological positions are evident. The first is the positivist paradigm, which can now be called neopositivist (Ebneyamini, 2022). The second option is the anti-positivist one called the interpretivist paradigm (Saunders and Bezzina, 2015). These two approaches are quite separate in the management sciences, and they are even considered as opposites (Sułkowski, 2012).

On the one hand, the positivist paradigm is characterized by quantitative research methods, while the interpretivist paradigm is characterized by qualitative methods (Saunders and Bezzina, 2015). In the case of the positivist approach, the value of the obtained knowledge is measured by the following characteristics:

- representativeness, understood as the degree to which reality has been studied and the possibility of drawing conclusions about the larger collective of objects studied (Blackler and Brown, 1983),
- intersubjectivity, i.e. the possibility of obtaining the same results in different research processes of the same objects (Frankfort-Nachmias and Nachmias, 2008),
- unambiguity (Ying He, 2021), which means the facts in organizational reality are understood in the same way by most researchers.

In the positivist approach, knowledge is obtained by verifying (or falsifying) hypotheses about relationships between variables (Ying He, 2021). However, it is necessary to keep in mind the neo-positivist asymmetry of the reliability of verification and falsification, introduced to the scientific world by Popper (2002).

It is still worth mentioning that if at the same time a normative approach is used in the field of ontology in the positivist approach, a special term is used in the literature for this kind of perception of organizational reality. This perception is called the science of construction management (in the original, "the science of construction management") (Runeson, 1997).

On the other hand, the interpretivist paradigm highlights the importance of qualitative methods in research in the organization sciences. Van Maanen wrote in the 1970s that qualitative research does not have a clear meaning in the management sciences and therefore it is a set of techniques for describing, interpreting, decoding or inquiring into the meaning but not on the frequency of certain phenomena occurring in the social world (Van Maanen, Sorenson and Mitchell, 2007).

A qualitative approach may seemingly mean describing a phenomenon to the exclusion of numbers or any statistics. However, this is not quite the case. It is more about a multi-disciplinary, multi-perspective approach based on a deep interpretation of phenomena (Cassell and Symon, 2006).

Taking into account both paradigms, it can be said that the avoidance of quantitative approaches causes difficulties in identifying the causes of phenomena, specifically organizational performance. This is also the root of the conflict between management science and so-called business practice. This is because, from the point of view of practitioners, the explanations and recommendations formulated are not very specific, impossible to implement, and they are sometimes simply based on contradictory assumptions (Cepeda and Martin, 2005).

Returning to the current division of research approaches in the management sciences into qualitative (positivist paradigm) and quantitative (interpretivist paradigm), the literature offers the view that there are rare examples of combining these two approaches (Sułkowski, 2013). This is for the following reason: If one assumes that a measure of the quality of scientific inquiry is high intersubjectivity, it can be shown that a contradiction is hidden in the interpretative approach. Interpretation is subjective, as the word interpretation itself indicates. The researcher subjectively apprehends the reality being studied (Gillespie and Cornish, 2010). The main differences between the positivist and interpretive approaches in management science are shown in Table 1.

Table 1: Metatheoretical assumptions in positivist and interpretive approaches

Elements	Positivist approach	Interpretive approach
Ontology	The researcher and the reality under study exist separately.	The researcher and the reality under study are inseparable.
Epistemology	Objective reality exists outside the human mind.	Knowledge about the world is created intentionally based on human experience.
Object of study	The object of study has immanent characteristics that exist independently of the researcher.	The object of study is an interpretation created in the mind of the researcher based on his experience.
Method	Statistics, content analysis	Phenomenology
Theory of truth	Correspondence theory of truth: testing the veracity of claims against reality.	Truth is intentional: an interpretation of the object of study within the researcher's experience.
Tractability	Relevance is ensured by properly collected data about reality.	Claims about reality are defensible by the researcher.
Reliability	Reproducibility of research results.	The researcher's awareness of the role of interpretation of reality.

Source: Cepeda and, Martin (2005)

According to the positivist approach, there is a real object, which is organizational reality (Saunders and Bezzina, 2015). Despite such a simple and internally consistent approach to the object under study, there is resistance to use a quantitative approach in management science. The opponents' argument is that the object of interest in management science is human beings, the motives of their behaviour, social relations, organizational culture, knowledge, conflict resolution, creativity, etc. This fact at the same time is an argument in support of the thesis that a qualitative approach is much better.

From the early 1980s onwards, some researchers began to question the position of positivist epistemology in management science. At that time, there was also a turn to interpretivist approaches (Nodoushani, 2000).

Summarizing the comparison of these two scientific approaches, it is worth pointing out that representatives of the qualitative approach accuse "positivists" of seeking the truth. Despite the assumptions of objectivity of scientific inquiry, in fact, the obstacle was the perception of reality in a subjective way. However, the same scientists propose to use an even more subjective research method - the interpretative qualitative approach (Malina, Nørreklit and Selto, 2011).

Representatives of the qualitative approach also point to epistemological problems arising from the use of the hypothetico-deductive model, which falls under the quantitative approach. This model involves deriving hypotheses from general scientific theories and then testing them with empirical data analysed quantitatively (!). It is believed that the models formulated in this trend are fragmentary, and the models themselves are priori constructs.

In the literature one can find opinions in favour of a quantitative view of organizational reality. Nodoushani (2000) even argues on the role of positivist epistemology in management science that if we abandon the positivist model of science in management science, then at the same time we must conclude that this world of organizations is something different from the world of nature. Once again, it is worth emphasizing that the division into qualitative and quantitative is not right.

Taking into consideration these opinions and answering the first research question (RQ1) which concerned the types of research methods which should be used to collect data on managerial work, it is possible to say that a

mixed method is the most appropriate to use (van Dun, Hicks and Celeste, 2015). This is one of the assumptions of the system of organizational terms which is a research methodology designed by the author to study the behaviours of team members in order to implement artificial management (Flak, 2020). This concept is briefly described in Section 3.

3. Theoretical Foundations of the TransistorsHead.com and Artificial Intelligence

The idea of replacing human managers with robots can be shaped by the idea of an intelligent knowledge system in an organization (Davern, 2007). In order to build an information system which could be used as research tools in implementing artificial management, the system of organizational terms was designed and applied to the system of online management tools (TransistorsHead.com) (Flak, 2023). The concept was being developed and tested during 2007-2018. The results of the research was published in many papers by the author (Alnajjar and Flak, 2016; Flak, 2018).

The view of managerial work has been changed since the birth of scientific management. At the beginning of the 20th century the picture of a manager was defined by his classical functions (set of activities), such as a planner, an organizer, a motivator and a controller. 50 years later, the idea of a manager has been dominated by two approaches and it has lasted until today.

Firstly, Koontz and O'Donneil (1964) launched a discussion on the meaning of managerial skills. A few years later an approach in which managerial work was represented by managerial skills was proposed. The managerial skill was than defined as an ability to work effectively as a team manager in order to build cooperative effort within the team (Katz, 1974). Secondly, in 1980 Mintzberg concluded that the managerial work in a team can be described in terms of 10 roles within interpersonal, informational and decisional areas which were common to the work of all types of managers. He defined a managerial role as an area of job activity which is undertaken by a manager.

These terms still do not recognize what a team manager really does (Sinar and Paese, 2016). So that, it is not possible to recognize (1) a sequence of managerial actions done one after another by a team manager, and (2) their content. These two aspects are included in the second research question (RQ2), concerning a way of implementing research methods into the online managerial tools to make the managerial work automated.

It is said that artificial intelligence has emerged as a promising and increasingly available technology for managerial decision-making. With the adoption of AI-enabled software, organizations can leverage various benefits of the technology, but they also have to consider the intended and unintended consequences of using the technology for managerial roles (Leyer and Schneider, 2021).

Regarding outcomes, teams perform equally as well under both types of management, with trends towards even superior performance from the AI-managed teams. Furthermore, team members perceive the interventions from both the AI and human manager as equally relevant (Gyory, et al, 2021). Some research on mixed human–robot teams show that teams that comprise both robots and humans has expanded attracting the attention of researchers from different disciplines (Wolf and Stock-Homburg, 2022). However, there are many knowledge gaps regarding the characteristics of human-AI teams, including a lack of understanding of how certain human-human teaming concepts may or may not apply to human-AI teams and how this composition affects team performance (McNeese, et al., 2021).

When one considers the use of artificial intelligence to replace a human manager with an artificial manager, the recent new technology ChatGPT cannot be overlooked. ChatGPT literature is still in its early stages, comparable to the formative years of "early childhood", where the understanding of ChatGPT's traits remains in a state of immaturity.

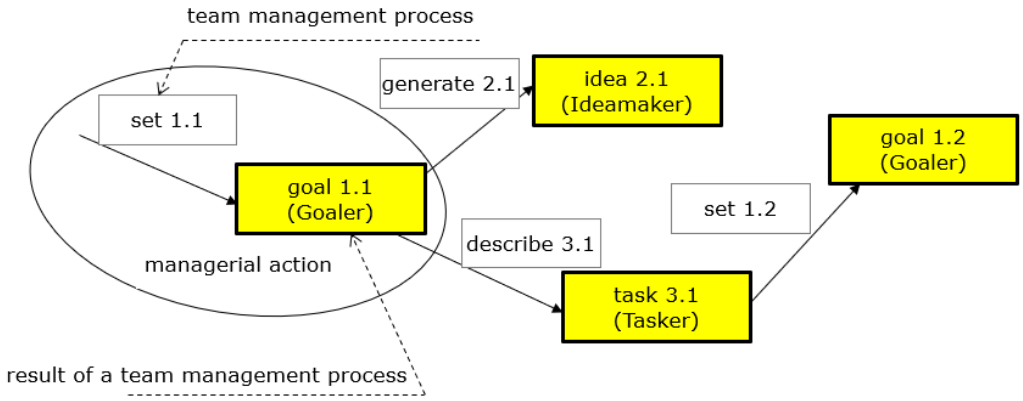
It is said, that there is a knowledge gap which the present qualitative enquiry aims to bridge, enlisting a diverse cohort of interviewees whose perspectives have contributed to a comprehensive philosophical framework for understanding the humanlike traits of ChatGPT in the role of a team manager (Lilly, Ismail, Abunaser, 2023). The point of contention of the current discussion is the Generative Pre-trained Transformer (GPT) models by the private company OpenAI. ChatGPT, a GPT-3.5-based application, has gathered much attention lately as of early 2023. ChatGPT can then be used to generate text by providing it with a prompt or a seed text, which it uses to generate a coherent and fluent response. GPT-3.5 and similar models can also perform other language-based tasks such as question answering, text completion and sentiment analysis. Thus, the model is, of course, not "thinking" but simply predicting what to say next (Burger, et.al., 2023).

However, ChatGPT has shown the potential of emerging general artificial intelligence capabilities, as it has demonstrated competent performance across many natural language processing tasks. Additionally, ChatGPT, as a type of generative AI, is a state-of-the-art LLM (Large Language Model) that is bundled with extensive capabilities which can enhance decision-making process, planning efficiency and AI-mediated communication fluency (Wong, Lian and Sun, 2023). What is more, ChatGPT has been one of the most talked about computer programs amongst management educators in recent weeks due to its transformative ability to change how assessments are undertaken and graded (Ratten and Jones, 2023).

Let's go back to the answer to the question of how to represent managerial work. As proved in the author's previous research, it can be found in the relationship between managerial roles and managerial skills. In order to play managerial roles a team manager should have some managerial skills. It results in the understanding of how playing managerial roles within their managerial skills by the day-to-day activities of managers', affect the managerial actions made by these managers. Therefore, the managerial action can be defined as a real activity, which a manager does in order to play a managerial role and have a certain managerial skill (Flak, Yang and Grzegorzek, 2017). The managerial action is a key term in the system of organizational terms.

The system of organizational terms is a methodological concept which consists of ontological and epistemological aspects designed for research team management aimed at artificial management. The philosophical foundation of the system of organizational terms is based on Wittgenstein's philosophy, his theory of facts (the only beings in the world) and "states of facts". According to this approach the organizational reality can be represented by events and things (Flak, 2020).

The system of organizational terms combines the resource approach and the process approach in the management science in a way that team management processes affect the team's resources. Therefore the managerial action is defined as a real activity, which a manager does in order to play a managerial role when he has a certain managerial skill (Yang, Flak and Grzegorzek., 2018), the managerial action structure consist of, e.g. event 1.1 and thing 1.1. It is shown in Figure 1.



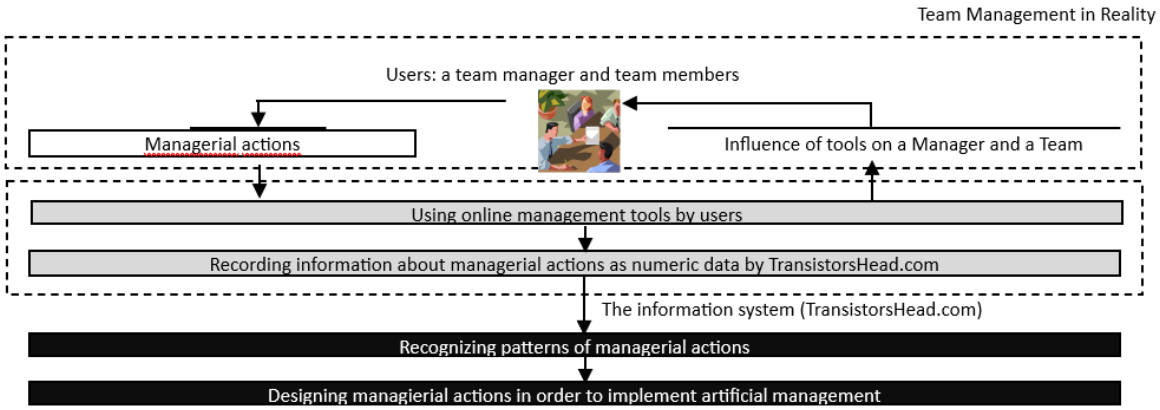
Source: Flak and Pyszka (2022)

Figure 1: The example of creating resources by processes in team management

As it is shown in Figure 1, when a team manager sets a goal (a team management process represented by event 1.1 – setting 1.1), it is possible to measure features of goal 1.1. If later (e.g. after describing a task – describing 1.1 and task 1.1) this team manager does the next setting of the same goal, he launches the next team management process. Then the features of this team management process are changed and represent the second version of this team management process (setting 1.2 and goal 1.2). The difference between features of goal 1.2 and goal 1.1. Let us do the reasoning on the team management process which happened in this period of time.

If it is only possible to measure things (e.g. goal 1.1), there should be a special research tool which could let us record the parameters of goal 1.1 keeping to the rule of minimum influence on a team manager. The solution is a research tool built into a management tool, which a team manager uses during day-to-day work. In other words, when a manager does something (e.g. set a goal) with a certain tool, this tool should record the parameters of the goal during the managerial work (e.g. before and after setting). This is the answer to the second research question (RQ2). The research method is built into the online management tools used by managers and their team members in their day-to-day work.

The Figure 2 shows a full course of the research methodology built into the information system included in online management tools and aimed at applying artificial management.



Source: Own elaboration

Figure 2: The use of the research methodology aimed at applying artificial management

4. Results of Research

The practical use of answers to both research questions (RQ1 and RQ2) as the results of field research are presented below. The information system TransistorsHead.com was used as a research tool to study behaviours of team managers and team members in order to apply artificial management. The results of the long-term observation is shown below. TransistorsHead.com recorded managerial actions of team management in 10 areas as follows: 1 - set goals (GOALS), 2 - describe tasks (TASKS), 3 - generate ideas (IDEAS), 4 - specify ideas (SPECIFICATIONS), 5 - create options (OPTIONS), 6 - choose options (DECISIONS), 7 - check motivation (MOTIVATION), 8 - solve conflicts (CONFLICTS), 9 - prepare meetings (MEETINGS), 10 - explain problems (PROBLEMS). The dashboard of the online management tools are presented in Figure 3.



Source: Flak (2023)

Figure 3: The dashboard of online management tools as research tools

The research was conducted from December 14, 2022 to January 14, 2023 among the students of the management faculty of Jan Kochanowski University of Kielce, Poland. The study involved 26 participants in 5 groups. Participants in the long-term observation were given the task of preparing the documentation for the program project on the YT channel in the Talent Show format. The task of the participants was as follows: formulate an organizing problem and solve this problem - propose a detailed format for the program and preparations for its implementation. Participants used the online management tools implemented in the information system TransistorsHead.com platform, including 10 online management tools, as shown in Figure 3 and described in the Section 3.

In this paper we show the example of one team, consisting of a team manager and 4 team members. Table 2 shows the total time spentusing the online management tools by each participant in this team. Table 3 contains the number of uses of each tool by individual team members.

For example, a team manager used a tool for setting goals (type 1) for 3792 seconds throughout the study (Table 2) and the goal-setting tool (type 1) was used by the team manager 141 times (Table 3). This does not mean that the manager produced 70 different goals. It means that he went back to the tool and worked on existing or new goals. It can be noted that team member 3 worked with some tools longer than the team manager, and also that team member 4 was hardly active at all while the team was working.

Table 2: Duration of a given managerial action taken by individual team members in seconds

User of online management tools	Type of managerial actions									
	1	2	3	4	5	6	7	8	9	10
team manager	3792	3092	3744	5670	1085	4916	3603	319	5605	5997
team member 1	365	5544	2567	2109	4644	786	1675	70	1957	4870
team member 2	696	1731	2826	1833	987	764	1837	405	1342	4943
team member 3	6550	2492	2814	1884	1215	652	1691	108	3285	3613
team member 4	0	0	375	0	0	0	0	0	1471	2395

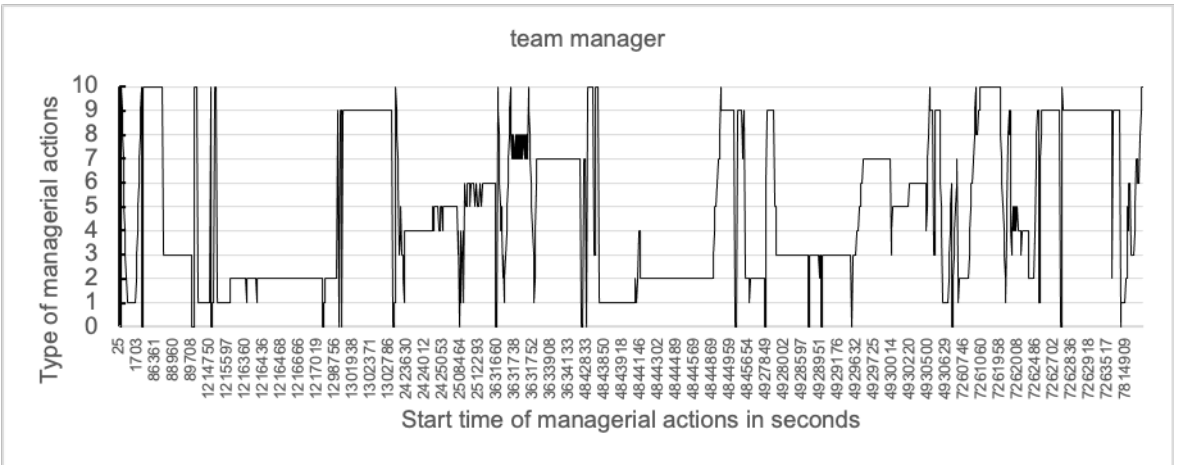
Source: Own elaboration

Table 3: Number of managerial actions taken by individual team members

User of online management tools	Type of managerial actions									
	1	2	3	4	5	6	7	8	9	10
team manager	141	324	179	96	96	84	138	32	254	93
team member 1	17	75	98	63	76	64	91	5	64	42
team member 2	13	31	97	53	71	71	89	10	31	38
team member 3	27	54	108	63	64	55	81	7	69	46
team member 4	0	0	30	0	0	0	0	0	44	6

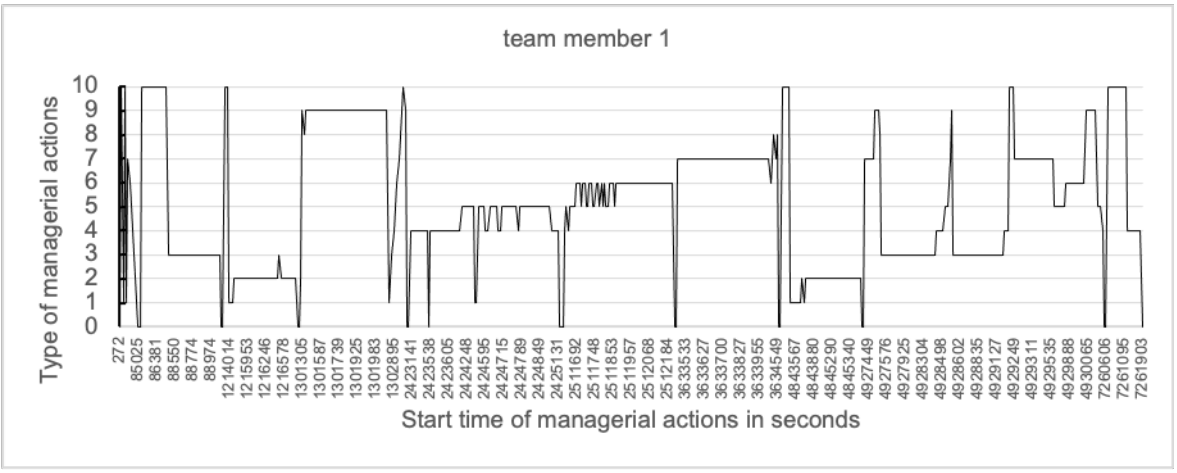
Source: Own elaboration

Figures 4, 5, 6, 7, 8 show the different trajectory of 10 managerial actions, recorded by 10 managerial tools in TransistorsHead.com in the team work period. The numbers in the types of managerial actions mean: 0 – no managerial action, 1 – setting goals, 2 – describing tasks, 3 – generating ideas, 4 – specifying ideas, 5 – creating option s, 6 – choosing options, 7 – checking motivation, 8 – solving conflicts, 9 – preparing meetings, 10 – explaining problems.



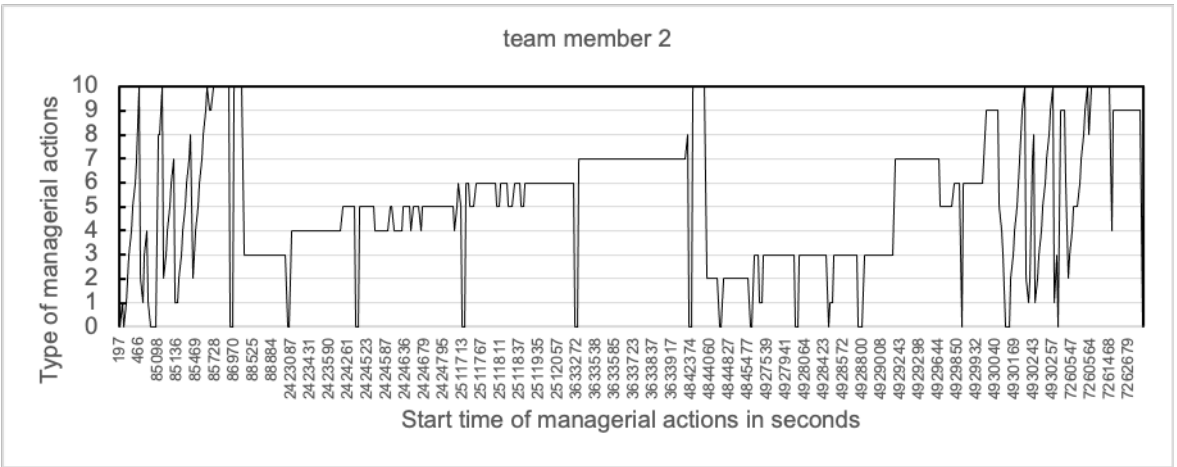
Source: Own elaboration

Figure 4: Trajectory of 10 managerial actions taken by a team manager



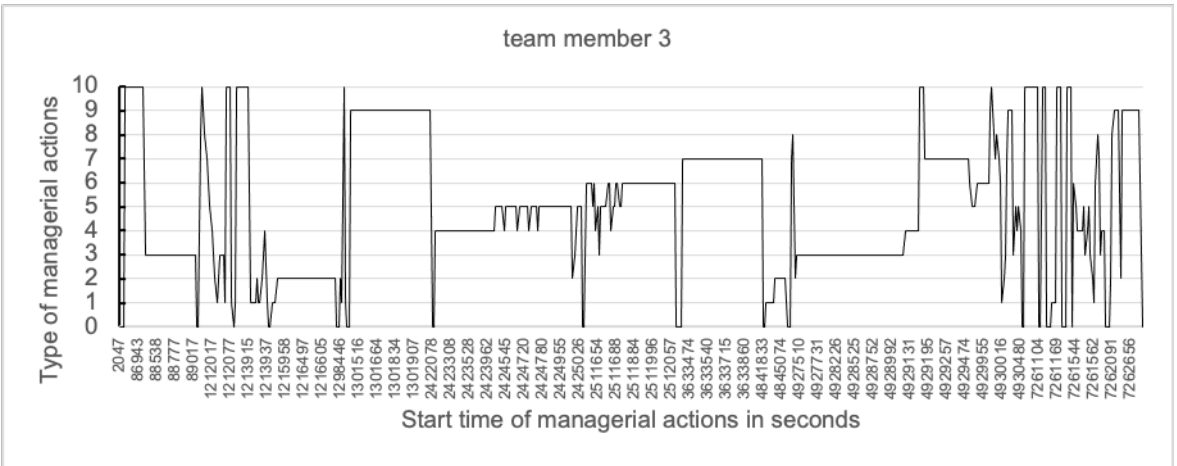
Source: Own elaboration

Figure 5: Trajectory of 10 managerial actions taken by a team member 1



Source: Own elaboration

Figure 6: Trajectory of 10 managerial actions taken by a team member 2



Source: Own elaboration

Figure 7: Trajectory of 10 managerial actions taken by a team member 3



Source: Own elaboration

Figure 8: Trajectory of 10 managerial actions taken by a team member 4

Interestingly, all team members' managerial actions had a different trajectory even when they were working on the same projects in the same team. This shows that the information system TransistorsHead.com can be used as a research tool to detect and collect data on human behaviour aimed at artificial management. Having recorded what a manager really does gives a necessary foundation to emulate his behaviour in team management, which is the practical use of answers to both research questions (RQ1 and RQ2).

5. Conclusion

The aim of the paper was to present a theoretical foundation for the information system which could be widely used by human managers in their day-to-day work and collect data on managerial work in order to apply artificial management. The content of this paper in Section 2 and 3 is the solution to the research problem which concerned the theoretical assumptions needed to design such an information system.

In Section 2, the answer to the first research question (RQ1) about types of research methods which should be used to collect data on managerial work was presented. The answer is the mixed method as the most appropriate method to study what a manager really does. The mixed method is understood as a combination of quantitative and qualitative methods.

Section 3 contains the answer to the second research question (RQ2) about a way of implementing research methods into the managerial tools. The answer is included in Figure 1, 2, and 3 as well as the theoretical and technical description.

In Section 4 there is an example of the practical use of the information system as a research tool which collects data on team management aimed at artificial management. Figures 4, 5, 6, 7, 8 show the trajectory of managerial actions which were taken by the team members. Having this set of data it is possible to emulate their behaviour which would lead to the automated managerial actions, as shown in Figure 2.

As the previous research of the authors proved that this approach can be used to apply artificial management (Yang, Flak and Grzegorzek, 2018; Flak and Pyszka, 2022), now ChatGPT shows the potential of emerging general artificial intelligence capabilities in team management automation. It has demonstrated competent performance across many natural language processing tasks (Amin, Cambria and Schuller, 2023). Additionally, ChatGPT, as a type of generative AI, is bundled with extensive capabilities which can enhance the decision-making process, planning efficiency and AI-mediated communication fluency (Wong, Lian and Sun, 2023). Therefore, the next step to implement artificial management seems to be a use of ChatGPT as an intelligent agent operating on the data collected by the TransistorsHead.com. This research approach is the next step in the development of replacing human managers with artificial management.

References

Al Lily, A.E., Ismail, A.F. and Abunaser, F.M., 2023. ChatGPT and the rise of semi-humans. *Humanities and Social Sciences Communications – Nature*, 10(626). https://doi.org/10.1057/s41599-023-02154-3

Alnajjar, J. and Flak, O., 2016. A managerial and linguistic perspective on researching manager behaviour aimed at replacing human managers with robots. *International Journal of Systems and Society*, 3 (2), pp. 14-29.

Amin, M.M., Cambria, E. and Schuller, B.W., 2023. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems*, 38, pp. 15-23. https://dx.doi.org/10.48550/arXiv.2303.03186

Blackler, F.H.M. and Brown, C.A., 1983. Qualitative research and paradigms of practice. *Journal of Management Studies*, 20(3), pp. 353-356. https://doi.org/10.1111/j.1467-6486.1983.tb00212.x

Bouschery, S.G., Blazevic, V. and Piller, F.T., 2023. Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2), pp. 139-153, https://doi.org/10.1111/jpim.12656

Burger, B., Kanbach, D.K., Kraus, S., Breier, M. and Corvello, V., 2023. On the use of AI-based tools like ChatGPT to support management research. *European Journal of Innovation Management*, 25(7), pp. 233-241. DOI 10.1108/EJIM-02-2023-0156

Cassell, C. and Symon, G., 2006. Taking qualitative methods in organization and management research seriously. *Qualitative Research in Organizations and Management: An International Journal*, 1(1), pp. 4-12. https://doi.org/10.1108/17465640610666606

Cepeda, G. and Martin, D., 2005. A review of case studies publishing in management decision 2003-2004. *Management Decisions*, 43(6), pp. 851-871. https://doi.org/10.1108/00251740510603600

Dash, R., McMurtrey, R. and Rebman, C., 2019. Application of artificial intelligence in automation of supply chain management. *Journal of Strategic Innovation & Sustainability*, 14(3), pp. 43-53. https://doi.org/10.33423/JSIS.V14I3.2105

Davenport, T.H. and Kirby, J., 2015. Beyond automation. *Harvard Business Review*, 6, https://hbr.org/2015/06/beyond-automation

Davern, M.J., 2007. Towards a unified theory of fit: Task, technology and individual. In: D. Hart and S. Gregor, eds. *Information systems foundations.* Acton: Australian National University E-Press. pp. 49-70. https://doi.org/10.22459/isftrr.11.2007.03

Ebneyamini, S., 2022. Towards developing a framework for conducting management studies using design research methodology. *International Journal of Qualitative Methods*, 21, pp. 1-7. https://doi.org/10.1177/16094069221112245

Ewenstein, B., Hancock, B. and Komm, A., 2016. Ahead of the curve: The future of performance management. *McKinsey Quarterly*, May, [online] Available at: http://www.mckinsey.com/business-functions/organization/our-insights/ahead-of-the-curve-the-future-of-performance-management, 16.05.2016

Flak, O., 2018. Teamwork research method based on the system of organizational terms and online management tools. *International Journal of Contemporary Management*, 17(2), pp. 7-34.

Flak, O., 2020. System of organizational terms as a methodological concept in replacing human managers with robots. *Lecture Notes in Networks and Systems*, Springer, pp. 471-500. https://doi.org/10.1007/978-3-030-12385-7_36

Flak, O., 2023. Ontology of online management tools aimed at artificial management implementation: An example of use in software design. In: H. Kaindl, M. Mannion and L. Maciaszek, eds. *Proceedings of the 18th International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE 2023.* pp. 621-628.

Flak, O., Yang, C. and Grzegorzek, M., 2017. Action sequence matching of team managers. In: M. De Marsico, G. Sanniti di Baja and A. Fred, eds. *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods ICPRAM*. pp. 386–393. https://dx.doi.org/10.5220/0006189203860393

Flak, O. and Pyszka, A., 2022. Evolution from human virtual teams to artificial virtual teams supported by artificial intelligence. Results of literature analysis and empirical research. *Management Issues*, 20 (96), pp. 48-69. https://doi.org/10.7172/1644-9584.96.3

Frankfort-Nachmias, C. and Nachmias, D., 2008. *Research methods in the social sciences*. 7th ed.New York: Worth.

Gillespie, A. and Cornish, F., 2010. Intersubjectivity: Towards a dialogical analysis. *Journal for the Theory of Social Behaviour*, 40(1), pp. 19-46. https://doi.org/10.1111/j.1468-5914.2009.00419.x

Gyory, J. T., Soria Zurita, N. F., Martin, J., Balon, C., McComb, C., Kotovsky, K., and Cagan, J., 2021. Human versus artificial intelligence: A data-driven approach to real-time process management during complex engineering design. *Journal of Mechanical Design*, 144 (2), 021405.

Katz, R.L., 1974. Skills of an effective administrator. *Harvard Business Review*, 52(5), pp. 90-102.

Keller, A., 2017. Challenges and directions in service management automation. *Journal of Network & Systems Management*, 25(4), pp. 884-901. https://doi.org/10.1007/s10922-017-9437-9

Khakurel, J. and Blomqvist, K., 2022. Artificial intelligence augmenting human teams. A systematic literature review on the opportunities and concerns. In: H. Degen and S. Natoa eds. *Artificial Intelligence in HCI*. Lecture Notes in Computer Science. 13336 Springer. pp. 51-68 https://doi.org/10.1007/978-3-031-05643-7_4

Koontz, H. and O'Donneil, C., 1964. *Principles of management*. New York: McGraw-Hill.

Leyer, M. and Schneider, S., 2021. Decision augmentation and automation with artificial intelligence: Threat or opportunity for managers? *Business Horizons*, 64(5), pp. 711-724. https://doi.org/10.1016/j.bushor.2021.02.026

Malina, M.A., Nørreklit, H.S.O. and Selto, F.H., 2011. Lessons learned: Advantages and disadvantages of mixed method research. *Qualitative Research in Accounting and Management*, 8(1), pp. 59-71. https://doi.org/10.1108/11766091111124702

McNeese, N. J., Schelble, B. G., Canonico, L. B., and Demir, M., 2021. Who/What is my teammate? Team composition considerations in human–AI teaming. *IEEE Transactions on Human-Machine Systems*, 51(4), pp. 288-299. https://doi.org/10.1177/21695067231192423

Mintzberg, H. (1980). The nature of managerial work. New York: Prentice-Hall.

Nodoushani, O., 2000. Epistemological foundations of management theory and research methodology. *Human Systems Management*, 19(1), pp. 71-80. http://dx.doi.org/10.3233/HSM-2000-19108

Popper, K.R., 2002. *The logic of scientific discovery*. London: Routledge.

Ratten, V. and Jones, P., 2023. Generative artificial intelligence (ChatGPT): Implications for management educators. *International Journal of Management Education*, 21(3), 100857. DOI:10.1016/j.ijme.2023.100857

Runeson, E., 1997. The role of theory in construction management research: Comment. *Construction Management Economics*, 15, pp. 299-302. https://doi.org/10.1080/014461997373033

Saunders, M. and Bezzina, F., 2015. Reflections on conceptions of research methodology among management academics. *European Management Journal*, 33(5), pp. 297-304, https://doi.org/10.1016/j.emj.2015.06.002

Sinar, E. and Paese, M., 2016. The new leader profile. *Training Magazine*, 46, pp. 46-50.

Sułkowski, Ł., 2012. Elements of organizational culture – theoretical and methodological problems. *Management*, 16(2), pp. 63–71. https://doi.org/10.2478/v10286-012-0056-y

Sułkowski, Ł., 2013. The identity of organization researchers. *Entrepreneurship and Management*, 14(1), pp. 56-66. http://dx.doi.org/10.2478/joim-2020-0049

Van Dun, D.H., Hicks, J.N. and Celeste, P.M., 2015. Values and behaviors of effective lean managers: Mixed-methods exploratory research. *European Management Journal*, 8, pp. 45-61. https://doi.org/10.1016/j.emj.2016.05.001

Van Maanen, J., Sorensen, J.B. and Mitchell, T.R., 2007. The interplay between theory and method. *Academy of Management Review*, 32(4), pp. 1145-1154, https://doi.org/10.5465/amr.2007.26586080

Wolf, F. D., and Stock-Homburg, R. M., 2022. How and when can robots be team members? Three decades of research on human–robot teams. *Group & Organization Management*, 48(6), pp. 1666-1744. https://doi.org/10.1177/10596011221076636

Wong, I.A., Lian, Q.L. and Sun, D.N., 2023. Autonomous travel decision-making: An early glimpse into ChatGPT and generative AI. *Journal of Hospitality and Tourism Management*, 56, pp. 253-263. https://doi.org/10.1016/j.jhtm.2023.06.022

Yanchinda, J., Yodmongkon, P., Pitipong, Ch., Chakpitak, N. and Goldsmith, P., 2011. Using ontologies for knowledge management. The Chaipattana Aerator Project. *Proceedings of the International Conference on Intellectual Capital, Knowledge Management & Organizational Learning.* pp. 806-817.

Yang, C., Flak, O. and Grzegorzek, M., 2018. Representation and matching of team managers: An experimental research. *IEEE Transactions on Computational Social Systems*, 5(2), pp. 311-323. https://dx.doi.org/10.1109/TCSS.2018.2812825

Ying, He, 2021. Revisiting Ellsberg's and Machina's paradoxes: A two-stage evaluation model under ambiguity. *Management Science*, 67(11), pp. 6897-6914. https://doi.org/10.1287/mnsc.2020.3835

Generative AI in Writing Workshops: A Path to AI Literacy

Sonja Gabriel
KPH Vienna/Krems, Vienna, Austria

sonja.gabriel@kphvie.ac.at

Abstract: The widespread use of generative AI tools which can support or even take over several part of the writing process has sparked many discussions about integrity, AI literacy and changes to academic writing processes. This paper explores the impact of generative artificial intelligence (AI) tools on the academic writing pro-cess, drawing on data from a writing workshop and interviews with university students of a university teacher college in Austria. Despite the widespread assumption that generative AI, such as ChatGPT, is widely used by students to support their academic tasks, initial findings suggest a notable gap in participants' experience and understanding of these technologies. This discrepancy highlights the critical need for AI literacy and underscores the importance of familiarity with the potential, challenges and risks associated with generative AI to ensure its ethical and effective use. Through reflective discussions and feedback from work-shop participants, this study shows a differentiated perspective on the role of generative AI in academic writing, illustrating its value in generating ideas and overcoming writer's block, as well as its limitations due to the indispensable nature of human involvement in critical writing tasks.

Keywords: Generative AI, Writing, AI literacy

1. Introduction

Since ChatGPT was released to the public in November 2022, the topic of generative artificial intelligence - especially those tools that can create, revise or translate texts - has been the subject of much controversy in higher education. The use of written, unsupervised work such as seminar papers, bachelor's and master's theses as an examination format has raised concerns (Song, 2024), as it is almost impossible to adequately assess independent work (e.g. by oral presentation of the written work) without investing much more time. As a result, various measures have been proposed, ranging from a complete ban on the use of generative AI to a laissez-faire approach that gives students complete freedom. However, it is important to note that neither of these extremes can be effective in the long term. The rapid technological progress in the field of generative AI means that these tools are already integrated into many standard applications, such as search engines or word processing software. On the other hand, students' unrestricted and unreflective use of these tools may widen the digital divide, as a lack of competence or access to generative AI tools could hinder successful study (Santiago-Ruiz, 2023). Between forbidding and allowing unregulated use of AI tools, there are various possibilities, as Buck and Limburg (2023) suggest. It is undeniable that writing processes will evolve with advancements. This will inevitably affect university education, where academic writing remains highly valued. To ensure successful and reflective writing development, it is necessary to consider these changes and initiate an adaptation of writer training (Limburg et al., 2023).

1.1 Generative AI and (Academic) Writing

When having a look at the stages of an academic writing process (Bailey, 2003), AI tools can be used in each one. Certain tools of generative AI can research literature, help learners to understand texts by enabling them to ask questions to the text, they can paraphrase and summarize, assist in structuring and brainstorming ideas as well as help with writing itself. Moreover, AI-tools can check spelling and grammar, do proof-reading and suggest changes as well as give feedback (if asked for). Research so far, quite often concentrates on using ChatGPT in many of the beforementioned tasks – for argumentative writing (Su et al., 2023), research (AlAfnan et al., 2023), grammar and style (Younis et al., 2023), retrieving references (Alyasiri et al., 2024) as a writing assistant (Imran & Almusharraf, 2023) or for foreign languages (Baskara, 2023). The use of ChatGPT in academic writing as a form of plagiarism was discussed in many papers as well (Jarrah et al., 2023). However, writers need to have certain skills – also called AI literacy – in order to use generative AI correctly and ethically.

The demand for AI literacy already existed before ChatGPT was introduced at the end of 2022. The competency framework by Long & Magerko (2020) comprises five competency areas with respective sub-areas, whereby the competencies are framed as questions: What is AI? What can AI do? How does AI work? How should AI be used? and finally, How do people perceive AI? Apart from this AI literacy framework, there have been many more developed since 2021 (Ehlers et al., 2023a). One recent framework based on qualitative and quantitative research is AIComp (Ehlers et al., 2023b). This model is structured in three competence areas that relate work, personal development and (social) environment, whereas these areas are subdivided into twelve competence

fields like creative problem-solving competence, critical thinking, ethical competence or system-design competence.

The TPACK Model which explains what teachers need to know to use technology for teaching and learning has been updated to integrate generative AI (Mishra et al., 2023): Technological Knowledge (TK) includes understanding the technology behind generative AI, for Technological Pedagogical Knowledge (TPK) teachers need to have an understanding how generative AI can support learning. Technological Content Knowledge (TCK) deals with the effects of generative AI on society, professions and every-day-life in general whereas Contextual knowledge (XK) takes systemic and cultural contexts into considerations. The models and approaches outlined above are very general, but provide a good basis for describing the skills required for (academic) writing. For example, knowledge of how generative AI works is a prerequisite for being able to recognize the advantages and limitations of individual tools and for using them competently and reflectively.

Regarding academic writing courses, Cardon et al. (2023) found that ChatGPT is mainly used to make writing processes more efficient, to generate ideas at the beginning of the writing process or to create first drafts. Moreover, generative AI is used less for the evaluation of written texts. As per the lecturers interviewed, the main points were about autonomy, authenticity and the ability to act. Technical expertise or background knowledge was not addressed in classes. In the first half of 2023 in particular, the topic of how ChatGPT can support academic writing processes was focused on in many research projects (Jen & Hj Salam, 2024) and discussed among (university) teachers. Tai et al. (2023) show how generative AI can improve grammar and text flow and thus is an enabler especially for linguistically weak students, but at the same time the authors point out that linguistic inequalities can occur among users.

1.2 Writing Literacy Supported by Generative AI

Rowland (2023) divides interaction with generative AI into nine different levels, which range between the two poles of "completely written by humans" and "completely generated by AI", although this approach does not judge which level is preferable (from a competence perspective). For example, level 7 describes students using generative AI as a critical "dialogue partner", while level 9 describes students adopting AI-generated texts unchanged. Another approach is taken by Limburg (2023). The use of generative AI is seen on a continuum of increasing competence and is located between the two poles of "avoiding learning" and "achieving better results". Depending on the extent and form in which generative AI tools are used, metacognitive competences are also considered, increasing when generative AI is used as a source of ideas or to obtain feedback.

Although the topic of academic writing has received a lot of attention in recent months, research mainly focus on how (university) teachers perceive the situation and how generative AI is integrated into courses (Cardon et al. 2023; Joshi et al., 2023) or whether and for what purposes students use AI tools (von Garrel et al, 2023; Cjeliebak et al., 2023). In terms of the writing process, there are studies that look at students' ethical attitudes towards the use of generative AI (Zorec et al., 2023), how teachers evaluate non-transparent and not clearly identifiable use of generative AI (Farazouli et al., 2023) or use of AI tools when writing in a foreign language (Lee et al, 2024). To date, there has been less focus on research into the teaching of skills related to the use of generative AI in the writing process in terms of metacognitive skills (Limburg, 2023). The present research project, carried out at an Austrian university teacher college, aims to partially fill this research gap by using a mixed-methods approach to investigate students' perspectives on the use of generative AI tools in the context of strengthening their AI literacy.

2. Research Design

Generative AI-tools have often been described as disruption (Alier et al., 2024) providing new possibilities for education but new challenges as well. The further development of teaching in terms of methodology and didactics at universities is therefore required being reflected in numerous handouts and guidelines in higher education (Tobor, 2024). New perspectives are opened up and at the same time there are complex challenges that cannot yet be assessed due to future changes of generative AI tools. For this reason, the following research questions are addressed:

- Which potentials and limitations or challenges do students identify when using generative AI tools in their (academic) writing process?
- To what extent does the writing process and / or writing motivation change due to the use of generative AI tools?

By answering these questions the research can contribute to the discussion of how writing skills can be developed using generative AI to assist. Moreover, the results will show the importance of AI-literacy and thus be able to influence curricular development when it comes to pre-service teacher education.

The research project consists of various different points of collecting data. The centerpiece of the project is an online workshop called Writing with generative AI-tools which started in October 2023 and took place four times in winter term and four times in summer term (cf. Fig. 1).

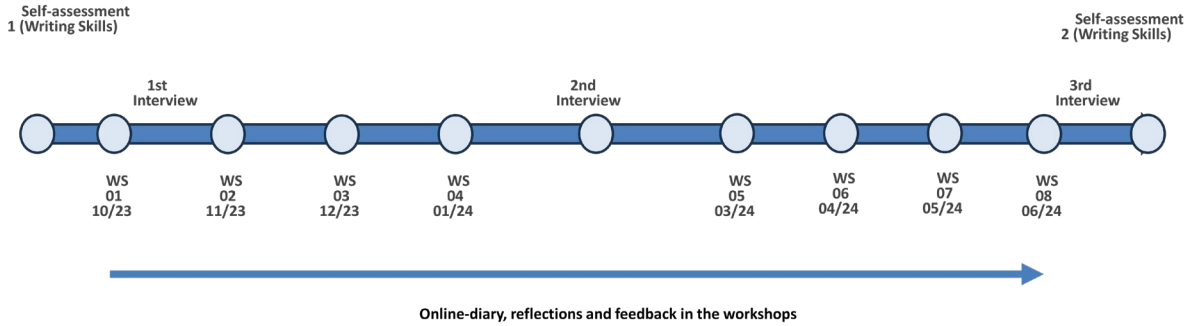


Figure 1: Research design (overview)

Each workshop lasts 60 minutes and is recorded for students who were not able to take part in the live sessions. The decision for a synchronous online session had to be taken because participants from different study programs and years registered. Each session focused on a different topic like planning and structuring texts with generative AI, paraphrasing, researching, revising texts. Each workshop included a short input at the beginning (for example how does text-generating AI work, short history of AI), was followed by discussing some tools (depending on the session's focus), their potentials and challenges and a short phase of trying out at least one of the tools. At the end, participants were asked to reflect on their experiences during the session.

The mixed methods research design consists of quantitative (online questionnaires at the beginning and the end of the project for self-assessment of writing skills) and different qualitative instruments (interviews, reflections and short written feedbacks). This combination ensures to get a rather complete understanding of the development of participants' writing skills. Data was collected from October 2023 to June 2024: Interviews which focus on writing skills, attitudes towards writing and AI as well as experience with writing and (generative) AI are planned at three different points of time (October 2023, February and June 2024). The interviews were conducted online, recorded, transcribed and analyzed using MAXQDA 24. The online questionnaires were made with MS forms and analyzed using MS Excel.

Additionally, participants were asked to keep an online diary for the time of the project which should be used to note down all their thoughts and reflections on using generative AI when writing. Only the participant him- or herself and the researchers have access to this diary.

After start of winter semester 2023/24 and more detailed information about dates and procedures, 18 students (17 female, one male) agreed to take part in the interviews and online diaries. These students were enrolled in three different degree programs in winter semester 23/24. The limitations of this research project are on the one hand the small size of the sample and on the other hand the fact that the writing workshop with AI tools is not integrated into the curriculum. Thus, the researchers depend on the voluntary participation and engagement of the participants. For this reason, only eight sessions of 60 minutes each were scheduled, which of course prevented an in-depth exchange with the participants in the sessions as well as a more detailed presentation and discussion of individual tools and their use. Due to overlaps in the students' timetables, it was also hardly possible to work with all participants at the same time during the synchronized sessions.

3. Selected Results

The following section will focus on first results from interviews (interview 1 and 2).

3.1 Knowledge About (Generative) AI

When using generative AI for writing, it is necessary to have at least some basic knowledge of AI and generative AI in order to recognize the potential and limitations of the tools. Therefore, one aim was to find out about the participants' prior knowledge. The 18 participants do not have much experience with generative AI, nor do they know exactly how to define AI and where to find it in everyday life. However, they assume that AI has an

important role to play: “I'm not that familiar with the topic, but when you get automated emails, for example, I wonder whether they might have been written using AI, or I once heard that some websites are also created with the help of AI.” (Interview 1_07, item 36) Some participants can give examples for the use of algorithms and AI but admit that they do not have enough knowledge about the topic: “Well, through all the algorithms, with advertising and social media and all these stories, of course. […] So you definitely come into contact with AI, probably a lot more often and I don't even know. It would be interesting to know.” (Interview 1_15, item 42)

When it comes to using generative AI, the experience is also very limited. Some of the participants have only heard of ChatGPT but not tried it before the workshops, some others have at least tried it, but none of them is a regular user. None of the interviewed students knows other tools than ChatGPT by name. How important it is to teach students at least some basic knowledge of (generative) AI can be seen from the answers by some participants in their interviews in February 2024. Learning about the history of AI, potentials and shortfalls of generative AI and their effects has made some participants think about it more carefully. “After each writing workshop, I reflected on the extent to which [AI] influences our future lives” (Interview 2_10, item 84).

3.2 First Experiences with Generative AI

After the first workshop which basically explained the structure of the research project, participants were asked to try out generative AI on their own without any further instructions. All respondents were provided with free and anonymized access to ChatGPT via the German education platform Fobizz. In the beginning of the next session (about a month later) they wrote down their experience. What can be seen from these encounters is that they were surprised in a positive way what generative AI is capable of but at the same time encountered some problems as well: “However, I notice that you have to phrase a "command" in more detail to get the expected results. Nevertheless, I have to say that I find it very helpful, especially as food for thought, an AI tool like ChatGPT is great. I would like to think more about AI in the future and get ideas for various projects/work etc.” (Participant 1, Session 2) They also noticed quickly that texts written by ChatGPT cannot simply be copied and used as they were: “[…] even though it always took some time to reorganize the texts. […] The result was a useful proposal that I could continue to work with.” (Participant 3, Session 2)

The second workshop introduced participants to prompting. The short input gave learners some advice on how to write better prompts for Large Language Models including examples. In the practical part of the workshop the participants were asked to improve a simple prompt they had written down before they got the input and to compare the results. Experience of the students show that prompting is (at least at the moment) important. “It helped me a lot when we learnt how to create the prompt,[…] which was very helpful for me when designing lesson plans, because then you have more ideas about how you could do it with the children. And if I didn't like the way it came out, then I changed something in the prompt and then you got a new answer, so I really enjoyed that.” (Interview 2_05, item 17)

3.3 Using Generative AI as a Discussion Partner

The Socratic Dialogue is an interactive method inspired by the classical Greek philosopher Socrates. It uses questioning and answering to stimulate critical thinking and generating ideas, diverging from traditional didactic teaching methods. This method is particularly valuable in teaching writing skills as it promotes critical analysis, requiring learners to construct coherent and persuasive arguments (Wilberding, 2021). This method helps to clarify thoughts and fosters exploration of diverse perspectives, enhancing students' ability to write from different viewpoints. Generative AI introduces a new dimension to the use of Socratic dialogues in education by simulating dialogues and engaging students in reflective learning as it generates questions that encourage deep analysis (Tabib & Alrabeei, 2024). Moreover, chatbots can provide immediate feedback on students' responses, guiding the learning process and suggesting areas for further exploration. Thus, it enables the creation of interactive writing exercises through written Socratic Dialogues, providing learners with practical experience in articulating their thoughts in writing.

The method of the Socratic Dialogue can be used with any Large Language Model. Learners only need to prepare the chatbot with a specific prompt. In Workshop 3, participants had a short introduction into the method of the Socratic Dialogue and were then asked to use a special prompt with ChatGPT 3.5 (cf. figure 2)

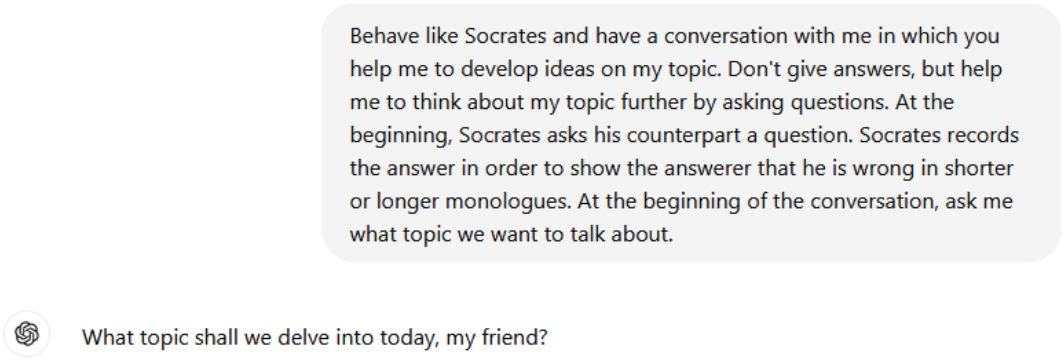


Figure 2: Prompt for starting Socratic dialogue

Learners were free on deciding which topic they would like to address and were then asked to have a chat for about 7 to 10 minutes and then write down their experience.

None of the students had known the method before and thus not used it with generative AI either. The reactions to the output were mostly positive as participants recognized the potential of "discussing" ideas and finding stronger arguments for or against a point of view: "Brilliant! Wow, this really opens up a new, practical and supportive method for dealing with my tasks in a structured way. I just asked a short question and received a variety of options as to which direction my thoughts and ideas could take." (Participant 5, Session 3)

However, some respondents also noticed that this method was not a way of getting an easy solution: "If the discussion goes on for a long time, you might get tired of always having your opinion reflected and looking at it in the mirror repeatedly, so to speak." (Participant 3, Session 3)

3.4 Influence on the Writing Process

The second round of interviews in February focused on the use of generative AI outside of the workshops and should show if there were already influences on the writing process. Most participants claimed that they did not have enough time to use the tools introduced during the workshops in detail as working with generative AI also needs some deeper thoughts in order to achieve better results. However, some areas can be identified which already at least slightly changed, especially when it comes to generating ideas or overcoming writer's block: "So if I'm typing in something, if I'm stuck somewhere and it gives me ideas where you think, OK, I haven't thought of that yet, yeah, I'll just go back to the library and look for a suitable book. So that's one of the motivations that AI has given me. Not so much for the writing itself, but just as a source of ideas." (Interview 2_10, item 65)

Generative AI is quite often seen as a virtual buddy that is willing to discuss every topic the students want to: "And then I asked the AI and then I sort of had a discussion with it. And of course I realize that there are also hallucinations, which we discussed last time. And I still have to do some research. But the fact that I get these approaches and these answers promptly via an AI tool is something I find very fascinating." (Interview 2_09, item 23) In some cases, participants also used AI tools to get different points of view: "There are other aspects in there that I would perhaps not write like that. And I think that's quite good." (Interview 2_7, item 52)

The range of tools which are regularly used has increased a lot. Most participants know exactly which AI tool to use for which task: "I also find DEEPL and DEEPL-Write particularly helpful, I have to be honest, because for me using academic language is rather difficult as I am not used to it." (Interview 2_01, item 11)

All these examples show that students use generative AI for tasks that would take a lot of time or put a high cognitive load on them. "So it's easier for me when I'm writing because I can reach my goal more quickly." (Interview 2_02, item 39) Also routine tasks like spell checking or summarizing texts are outsourced. However, what participants also learned is that AI cannot take all the work when dealing with writing tasks. "Yeah, otherwise I'm a bit ambivalent about writing or improving texts or something like that, because it's still like that, first of all you have to read through everything very often and then you actually want to rewrite it yourself. […] but I wouldn't just copy it without changing." (Interview 2_08, item 16)

Even when some respondents answered that they do not see much change in their writing process when directly asked, their answers show that there have been at least slightly different approaches towards collecting ideas, researching, starting to write, rephrasing and revising their first draft of a paper.

4. Conclusion

First results of the writing workshop with AI tools show that participants did not have a lot of experience and knowledge regarding AI and generative AI although many surveys [] claim that university students use ChatGPT & Co quite regularly to help them with their tasks. The danger of over or under estimating generative AI is quite high when there is not at least basic knowledge. The relevance of AI literacy with regards to potential, challenges and even risks of these tools is essential for targeted and ethically correct use as the results of the interviews and reflections during the workshops show.

Without information and advice about, for example, the risk of hallucinations caused by large language models, participants would have accepted the results without reflection, whereas reflections and interviews after the first workshops showed that students are now paying more attention to checking texts. The interviewees see great potential in the fact that generative tools can take some of the pressure off writing, or relieve them of routine tasks, without them having to worry that they are not doing enough themselves.

The second round of interviews provides a mixed picture of the role of generative AI in the academic writing process, highlighting both its value and limitations. While participants have not used the tools extensively due to time constraints and the depth of engagement required, they acknowledge the benefits of AI in generating ideas and overcoming writer's block, seeing it as a source of inspiration and a means of expanding perspectives rather than a direct writing aid. The interviews reveal an increase in the specificity and variety of generative AI tools used, tailored to tasks such as academic language use, text correction and rephrasing, suggesting that students are becoming increasingly adept at using AI for specific writing challenges. However, there's a clear consensus about the importance of personal involvement in critical writing tasks, with AI seen as a supportive tool that facilitates certain aspects of the writing process but cannot replace the critical engagement required to produce sound texts. This reflects a balanced acceptance of generative AI, where it is valued for its ability to support and enhance the writing process without diminishing the essential role of human creativity and judgement.

As a first conclusion, it can be stated that students – although they might already use generative AI in their writing process need to be accompanied in order to acquire necessary skills. The integration of generative AI into academic writing processes requires a thoughtful, informed approach that balances technological utility with critical human oversight. The development and dissemination of comprehensive AI literacy programs is crucial to empower students to navigate the complexities of generative AI tools, and to enable them to use these technologies responsibly and effectively. In addition, the curriculum must evolve to include the ethical considerations and practical skills associated with AI, ensuring that students are equipped to critically evaluate and integrate AI-generated content into their work in ways that enhance their integrity and creativity.

References

A. Younis, H., Osamah Mohammed, Muthmainnah, Sahib, T. M., Akhtom, D., Hayder, I. M., Salisu, S., & Shahid, M. (2023). ChatGPT Evaluation: Can It Replace Grammarly and Quillbot Tools? British Journal of Applied Linguistics, 3(2), pp. 34–46.

AlAfnan, M. A., Samira Dishari, Marina Jovic, & Koba Lomidze. (2023). ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses. Journal of Artificial Intelligence and Technology.

Alier, M., García-Peñalvo, F.-J., & Camba, J. D. (2024). Generative Artificial Intelligence in Education: From Deceptive to Disruptive. *International Journal of Interactive Multimedia and Artificial Intelligence*, *8*(5), 5.

Alyasiri, O. M., Salman, A. M., Akhtom, D., & Salisu, S. (2024). ChatGPT revisited: Using ChatGPT-4 for finding references and editing language in medical scientific articles. Journal of Stomatology, Oral and Maxillofacial Surgery, 101842

Bailey, S. (2003). Academic writing: A practical guide for students. Nelson Thornes.

Baskara, F. R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. International Journal of Education and Learning, 5(1), pp. 44–55.

Buck, I., & Limburg, A. (2023). Hochschulbildung vor dem Hintergrund von Natural Language Processing (KI-Schreibtools). Ein Framework für eine zukunftsfähige Lehr- und Prüfungspraxis. die hochschullehre, 9, pp. 70–84.

Cardon, P., Fleischmann, C., Aritz, J., Logemann, M., & Heidewald, J. (2023). The Challenges and Opportunities of AI-Assisted Writing: Developing AI Literacy for the AI Age. *Business and Professional Communication Quarterly*, *86*(3), pp. 257–295.

Cardon, P., Fleischmann, C., Aritz, J., Logemann, M., & Heidewald, J. (2023). The Challenges and Opportunities of AI-Assisted Writing: Developing AI Literacy for the AI Age. *Business and Professional Communication Quarterly*, *86*(3), pp. 257–295.

Cieliebak, M., Drewek, A., Jakob Grob, K., Kruse, O., Mlynchyk, K., Rapp, C., & Waller, G. (2023). *Generative KI beim Verfassen von Bachelorarbeiten: Ergebnisse einer Studierendenbefragung im Juli 2023*.

Ehlers, U.-D., Lindner, M., & Rauch, E. (2023). AI-Comp—Future Skills for a world shaped by AI. https://next-education.org/downloads/2023-12-14-AIComp-FutureSkills-Modelleng.pdf

Ehlers, U.-D., Lindner, M., Sommer, S., & Rauch, E. (2023). AICOMP - Future Skills in a World Increasingly Shaped By AI. Ubiquity Proceedings.

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2023). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, pp. 1–13.

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. Contemporary Educational Technology, 15(4), ep464.

Jarrah, A. M., Wardat, Y., & Fidalgo, P. (2023). Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say? Online Journal of Communication and Media Technologies, 13(4).

Jen, S. L., & Hj Salam, A. R. (2024). A Systematic Review on The Use of Artificial Intelligence in Writing. International Journal of Academic Research in Progressive Education and Development, 13(1), pp. 1819-1829.

Joshi, I., Budhiraja, R., Tanna, P. D., Jain, L., Deshpande, M., Srivastava, A., Rallapalli, S., Akolekar, H. D., Challa, J. S., & Kumar, D. (2023). *"With Great Power Comes Great Responsibility!": Student and Instructor Perspectives on the influence of LLMs on Undergraduate Engineering Education*.

Lee, Y.-J., Davis, R. O., & Lee, S. O. (2024). University students' perceptions of artificial intelligence-based tools for English writing courses. *Online Journal of Communication and Media Technologies*, *14*(1), e202412.

Limburg, A. (2023, July 4). *Akademisch Schreiben und Lernen mit KI - Empfehlungen für eine zukunftsfähige Lehre*. KI in Studium und Lehre, Mainz. https://lehre.uni-mainz.de/ki-eventreihe-jgu/#collapse4

Limburg, A., Bohle-Jurak, U., Buck, I., Grieshammer, E., Gröpler, J., Knorr, D., Mundorf, M., Schindler, K., & Wilder, N. (2023). Zehn Thesen zur Zukunft des Schreibens in der Wissenschaft (Diskussionspapier Nr. 23). Hochschulforum Digitalisierung.

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16.

Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the age of ChatGPT and Generative AI. *Journal of Digital Learning in Teacher Education*, *39*(4), pp. 235–251.

Rowland, D. R. (2023). *A conceptual framework for discussing the human-AI writing continuum*.

Santiago-Ruiz, E. (2023). Writing with ChatGPT in a context of educational inequality and digital divide. International Journal of Education & Development Using Information & Communication Technology, 19(3), pp. 28–38.

Song, N. (2024). Higher education crisis: Academic misconduct with generative AI. Journal of Contingencies and Crisis Management, 32(1), e12532.

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. Assessing Writing, 57, 100752.

Tabib, F. M., & Alrabeei, M. M. (2024). Can Guided ChatGPT Use Enhance Students' Cognitive and Metacognitive Skills? In A. Al-Marzouqi, S. A. Salloum, M. Al-Saidat, A. Aburayya, & B. Gupta (Eds.), Artificial Intelligence in Education: The Power and Dangers of ChatGPT in the Classroom (Vol. 144, pp. 143–154). Springer Nature Switzerland.

Tai, A. M. Y., Meyer, M., Varidel, M., Prodan, A., Vogel, M., Iorfino, F., & Krausz, R. M. (2023). Exploring the potential and limitations of ChatGPT for academic peer-reviewed writing: Addressing linguistic injustice and ehtical concerns. *Journal of Academic Language & Learning*, *17*(1), T16–30.

Tobor, J. (2024). *Blickpunkt—Leitlinien zum Umgang mit generativer KI*. Hochschulforum Digitalisierung. https://hochschulforumdigitalisierung.de/news/blickpunkt-ki-leitlinien/

von Garrel, J., Mayer, J., & Mühlfeld, M. (2023). *Künstliche Intelligenz im Studium Eine quantitative Befragung von Studierenden zur Nutzung von ChatGPT & Co*.

Wilberding, E. (2021). Socratic Methods in the Classroom: Encouraging Critical Thinking and Problem Solving Through Dialogue (1st ed.). Routledge.

Zorec, M.-C., Granegger, N., & Gorden, R. (2023). Wer hat das letzte Wort? Ethischer Wertekompass von Student*innen zur Verwendung von CGPT im Hochschulkontext. *zeitschrift für interdisziplinäre schreibforschung*, pp. 41-70.

Generative AI and Educational (In)Equity

Sonja Gabriel
University College of Christian Churches of Teacher Education, Vienna, Lower Austria

sonja.gabriel@kphvie.ac.at

Abstract: This paper examines the complex relationship between generative artificial intelligence (AI) and educational equity, analysing both the opportunities and challenges presented by these emerging technologies in educational contexts. The paper begins by establishing fundamental distinctions between educational equality and equity, emphasizing how various socioeconomic, cultural, and systemic factors contribute to persistent educational disparities. It then provides a comprehensive overview of generative AI technologies, particularly focusing on Large Language Models (LLMs) and their applications in educational settings. The analysis reveals several promising applications of generative AI for promoting educational equity, including enhanced accessibility features for students with disabilities, personalized learning experiences, and the creation of Open Educational Resources (OER). The paper highlights how AI-assisted tutoring, incorporating Socratic dialogue methods, and AI-generated feedback systems can provide valuable educational support, especially in resource-constrained environments. These technologies demonstrate potential in breaking down traditional barriers to education by offering multilingual support, adaptive learning materials, and immediate feedback mechanisms. However, the paper also addresses significant challenges and risks associated with implementing generative AI in education. These include concerns about digital divides, both in terms of access to technology and digital literacy skills, as well as the potential for AI systems to perpetuate existing biases. The research emphasizes the importance of thoughtful integration of AI technologies in educational settings, suggesting that the most effective approach may be a balanced combination of human instruction and AI-supported learning. By examining these various aspects, the paper contributes to ongoing discussions about how to harness generative AI's potential while ensuring its implementation promotes, rather than hinders, educational equity. The findings have significant implications for educators, policymakers, and educational institutions working to create more equitable learning environments in an increasingly technology-driven world.

Keywords: Generative AI, Education, Equity, Potential and challenges

1. Equality and Equity in Education

Educational equality and equity are complex, multifaceted issues with significant implications for individuals and society as a whole. Educational equality refers to the principle that all students should have equal access to high-quality educational opportunities, regardless of their socioeconomic, cultural, racial, or other background (Eurydice, 2020). Equity, on the other hand, recognizes that students have different needs and backgrounds and aims to provide each learner with the specific resources and support they need to achieve equal outcomes. The history of educational equality and equity dates back to ancient civilizations, but the modern understanding of these concepts has evolved significantly over time. In the 20th century, landmark events such as the Universal Declaration of Human Rights (1948) and the Convention against Discrimination in Education (1960) established education as a fundamental human right and emphasized the importance of equal access to education for all.

In the context of education, the distinction between equality and equity can be stated in the following ways:

- Equality would ensure that each student is provided with the same number of books. Equity, however, would guarantee that each student has access to the books they require, even if it means that some students receive more books than others.
- Equality would mean that all students are taught in the same way. Equity, on the other hand, would employ a variety of teaching methods to accommodate different learning styles and needs.
- Equality would result in all students being assessed using the same test. Equity, however, would involve the use of different assessment methods to account for diverse strengths and challenges.

However, despite progress in expanding educational opportunities, numerous factors contribute to persistent inequity in educational systems worldwide. Socioeconomic status is one of the most significant factors, with students from low-income families often facing disadvantages such as limited access to resources, lower-quality schools, and less parental involvement. Geographic location also plays a crucial role, with urban-rural divides and differences between affluent and economically disadvantaged neighbourhoods resulting in disparities in school funding, teacher quality, and available resources.

Race, ethnicity, and gender continue to be significant factors in educational inequality, particularly in countries with histories of systemic discrimination. Students with disabilities, and/or limited access to technology also face significant barriers to equal educational opportunities. Language barriers represent an additional factor

contributing to educational inequality, particularly for immigrant students or linguistic minorities. A lack of proficiency in the language of instruction can hinder academic progress and social integration. Research has consistently shown persistent achievement gaps between different racial, ethnic, and socioeconomic groups, highlighting the ongoing challenges in achieving educational equity. The COVID-19 pandemic has further increased these inequalities, with the shift to remote learning highlighting stark differences in access to reliable internet, devices, and supportive home learning environments. Students from disadvantaged backgrounds have been disproportionately affected by learning losses and disengagement during the pandemic, widening pre-existing achievement gaps. In this context, the emergence of generative AI in education presents both opportunities and challenges for advancing educational equity. While these technologies have the potential to personalise learning, improve accessibility, and democratise access to educational resources, they also risk widening existing digital divides and reinforcing biases if not implemented responsibly.

2. Generative AI in Education: An Overview

Generative AI refers to artificial intelligence systems capable of creating new content, including text, images, audio, and video, based on patterns and information learned from large datasets. These systems use sophisticated machine learning techniques, particularly deep learning and neural networks, to analyse and understand complex patterns in data, enabling them to mimic human-generated content. At the core of many generative AI systems are large language models (LLMs) or diffusion models trained on vast corpora of text or visual data. These models learn to predict patterns and relationships within the data, allowing them to generate coherent and contextually relevant content in response to prompts or input. Some of the most prominent examples of LLMs include OpenAI's GPT series, Google's Gemini, and Microsoft's Copilot. These models have demonstrated remarkable capabilities in generating human-like text, answering questions, and even engaging in open-ended conversations. Multimodal generative AI systems, such as DALL-E, Stable Diffusion, and Midjourney, combine language and visual understanding to generate or manipulate images based on textual descriptions. These systems use diffusion models trained on large datasets of image-text pairs to learn the relationships between visual and textual representations. For a more detailed explanation of how generative AI works, refer to Dhamani and Engler (2024).

In educational contexts, generative AI might transform teaching and learning experiences. From offering personalized learning opportunities to supporting rapid content creation, these technologies have the potential to enhance educational practices in meaningful ways. Additionally, generative AI can provide valuable language support, improve accessibility for diverse learners, and offer immediate feedback on student performance. Each of these benefits contributes to the evolving role of AI in education, with the promise of making learning more adaptable, inclusive, and efficient (Holmes et al., 2019). However, the emergence of generative AI in education also raises concerns about increasing existing digital divides and potentially widening educational inequalities. The first digital divide, characterized by unequal access to technology and internet connectivity, remains a significant barrier in many regions. The second digital divide, which focuses on disparities in digital skills and usage patterns, will become more significant with generative AI, as the ability to effectively develop and interact with AI becomes increasingly important. Furthermore, the emergence of a zero digital divide highlights pre-existing socio-economic factors that influence technology adoption and effective use, even when access is available (Holmes et al., 2019). Students from disadvantaged backgrounds may lack the necessary support, resources, or cultural capital to fully benefit from AI-enhanced learning opportunities. As generative AI continues to develop and enter educational contexts, it is crucial to consider both the promises and challenges it presents for educational equity. The following sections explore these issues in more detail, looking at how generative AI can be used to promote inclusivity and accessibility, while also addressing the risks and potential unintended consequences.

3. The Promise of Generative AI for Educational Equity

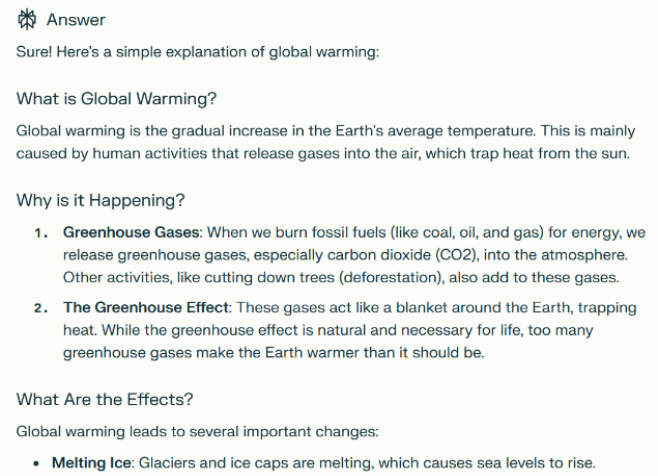
In the constantly evolving landscape of education, meeting the diverse needs of learners has become both a key challenge and a critical opportunity for fostering equitable and effective learning environments. As educational institutions grapple with increasingly heterogeneous student populations, the need to provide personalised, inclusive and adaptive learning experiences has never been more important. This shift towards learner-centred approaches recognises that each student brings a unique set of skills, backgrounds and learning strategies to the classroom, requiring a multi-faceted strategy to address these diverse needs.

Generative AI technologies offer promising tools to support this learner-centred paradigm. Using advanced natural language processing, computer vision and machine learning techniques, generative AI can create customised learning content, provide targeted feedback and adapt to individual learners' needs. From generating personalised explanations and practice problems to creating engaging visual aids and interactive elements, generative AI has the potential to improve the way we approach education and support diverse learners.

3.1 Accessibility and Inclusion

When considering the advantages of generative AI in education, Large Language Models (LLMs) are particularly noteworthy for their ability to create customized learning materials. Although LLMs are not yet capable of real-time adaptation to individual students, they can generate a wide range of practice problems and examples on demand. This feature allows students to gain more practice in areas where they face difficulties. Additionally, LLMs enable quick production of supplementary materials, such as study guides or summaries of complex texts, making it easier for teachers to provide a broader selection of differentiated learning resources without significant time investment. Moreover, LLMs can contextualize learning materials based on a student's interests or relevant real-world applications. For instance, a physics lesson could be framed using sports analogies for an athletic student or music-related examples for a student interested in sound and acoustics. Similarly, mathematical concepts could be explained through video game scenarios for a gaming enthusiast or cooking analogies for a culinary-minded learner. This level of contextualization makes learning more engaging and meaningful, supporting the constructivist principle that learners build new knowledge from their personal interests and experiences. While personalized learning enables tailored educational experiences based on individual needs, accessibility ensures that these personalized opportunities are available to all learners, regardless of their abilities or circumstances, highlighting how generative AI can address both customization and inclusivity in education.

Advanced generative AI models are also capable of converting written text into natural-sounding speech and vice versa, significantly benefiting learners with visual impairments or reading difficulties. For example, a student with dyslexia could use AI to convert textbook content into audio format, making it easier to understand. Similarly, students with hearing impairments could leverage speech-to-text capabilities to participate more actively in lectures or discussions. Generative AI can produce detailed textual descriptions of images, graphs, and diagrams for students with visual impairments, providing richer contextual understanding. This ability goes beyond basic recognition, as AI can create nuanced, context-aware descriptions that capture subtle details and relationships within visual content. Auto-generated subtitles for videos—such as those found on YouTube—are another feature of generative AI that can assist both visually impaired students and those accessing content in foreign languages. These subtitles, available in numerous languages, can enhance learning comprehension, reduce cognitive load (the mental effort required for learning and problem-solving), and increase overall satisfaction. Some AI apps go further, offering users the ability to ask questions about video content or generate summaries. For students with learning disabilities, such as dyslexia or dyscalculia, generative AI can provide personalized learning experiences by tailoring educational content to their needs. AI systems can generate explanations, examples, and practice problems at varying levels of complexity.


Answer

Sure! Here's a simple explanation of global warming:

What is Global Warming?

Global warming is the gradual increase in the Earth's average temperature. This is mainly caused by human activities that release gases into the air, which trap heat from the sun.

Why is it Happening?

1. **Greenhouse Gases**: When we burn fossil fuels (like coal, oil, and gas) for energy, we release greenhouse gases, especially carbon dioxide (CO2), into the atmosphere. Other activities, like cutting down trees (deforestation), also add to these gases.
2. **The Greenhouse Effect**: These gases act like a blanket around the Earth, trapping heat. While the greenhouse effect is natural and necessary for life, too many greenhouse gases make the Earth warmer than it should be.

What Are the Effects?

Global warming leads to several important changes:

- **Melting Ice**: Glaciers and ice caps are melting, which causes sea levels to rise.

Figure 1: Explaining a complex concept in simple words using perplexity.ai, prompted on 30 July 2024

For instance, a generative AI tool could rephrase complex texts into simpler language or offer multiple explanations of a concept using different analogies or contexts (see Figure 1). This adaptability helps students progress at their own pace, working through material as they build up to more challenging content.

Learners with mobility impairments may also benefit from generative AI's ability to create content from minimal input. For example, AI can expand brief voice commands or short text entries into full sentences or paragraphs, significantly reducing the physical effort required for writing tasks. Moreover, LLMs can simplify complex text without losing the core meaning, a valuable feature for students with cognitive disabilities, non-native speakers, or those who struggle with advanced vocabulary. Generative AI can also generate multiple versions of the same content at different levels of complexity, allowing students to gradually tackle more demanding material. Additionally, AI-generated images can help reduce learning barriers by creating visual aids from text descriptions (see Figure 2), assisting visual learners or those who find text-heavy material challenging. For instance, a description of a historical event could be transformed into a simple illustration to aid comprehension.

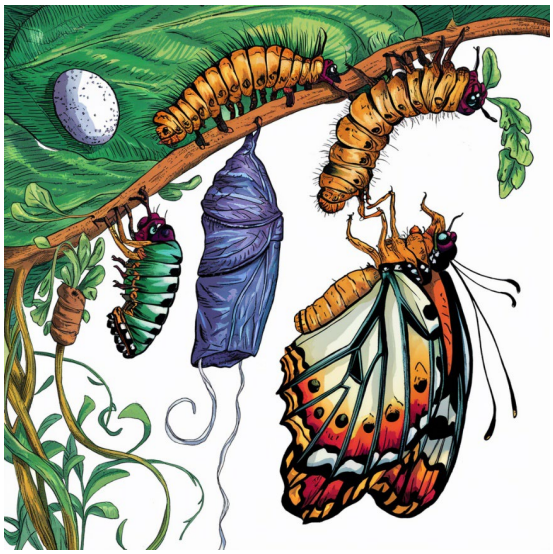


Figure 2: Illustrate the development of a butterfly (AI generated with ideogram.ai, prompted on 7 July 2024)

When it comes to language learning, particularly for students whose first language is not the language of instruction, LLMs are a valuable tool. Generative AI can quickly provide translations or explanations of unfamiliar terms, helping bridge language gaps and making content more accessible to multilingual learners.

3.2 Open Educational Resources and Practices

Open Educational Resources (OER) and Open Educational Practices (OEP) have significantly advanced the democratization of education. OER includes freely accessible and adaptable teaching, learning, and research materials available in various formats under open licenses (for more details, see UNESCO 2019). By reducing or eliminating the financial burden of educational materials, OER make quality learning resources more accessible to learners from diverse socio-economic backgrounds. The adaptability of OER also allows educators to localize and customize content to meet the needs of different learners, making education more inclusive. This flexibility is particularly valuable for marginalized communities and learners with disabilities, who are often underserved by traditional educational resources. OEP refers to the practices surrounding the creation, use, and management of these open resources with the aim of improving educational quality and fostering innovation. These practices encourage the integration of multiple perspectives and experiences into the educational process, leading to more culturally responsive and representative curricula. By making high-quality educational resources and innovative teaching practices widely available, they contribute to leveling the playing field between well-resourced and under-resourced educational institutions.

The advent of generative AI adds a new dimension to both OER and OEP: Generative AI can accelerate the production and customization of OER, enabling the rapid creation of diverse and culturally relevant materials. One significant advantage is that content produced by large language models is not copyrighted, allowing it to be freely distributed to learners worldwide. Similarly, AI can enhance OEP by providing intelligent support to educators in the design and implementation of open practices. However, the integration of generative AI in this context raises concerns about equitable access and digital literacy. While AI presents opportunities for

personalized learning and expanded access to resources, it also risks exacerbating existing digital divides, particularly in terms of AI literacy and access to advanced technologies. As the educational landscape continues to evolve with these technological advances, it is crucial to ensure that the principles of openness and equity that underpin OER and OEP are not only preserved but enhanced by the inclusion of generative AI.

3.3 AI-Assisted Tutoring

Large Language Models (LLMs) can be utilized to answer students' questions when they are stuck or need further explanation. However, they can also complete tasks on behalf of students, which could prevent learners from developing the necessary skills. To mitigate the risk of students becoming overly reliant on AI chatbots, there are projects like Khanmigo (https://www.khanmigo.ai). Instead of providing direct answers, Khanmigo guides students through questions, helping them arrive at the correct solution on their own. While this system is a commercial product (available only in the U.S. as of July 2024), there are also customized GPTs offered by individuals globally, based on ChatGPT-4.

The concept of Socratic dialogue, rooted in ancient Greek philosophy, offers a valuable pedagogical approach that involves systematic questioning and critical inquiry. Saran and Neisser (2004) provide a detailed explanation of this method, which is designed to stimulate critical thinking, challenge assumptions, and lead to the discovery of new knowledge. Key elements of Socratic dialogue include elenchus (testing beliefs), aporia (a state of puzzlement when contradictions are revealed), and maieutics (the 'midwifing' of ideas). In educational contexts, the Socratic method is an effective teaching strategy that fosters active learning, critical thinking, and metacognitive awareness. By encouraging students to question their assumptions and analyze arguments, the method promotes conceptual understanding, aligning well with constructivist learning theories. Furthermore, it enhances metacognitive skills, develops communication abilities, and cultivates intellectual humility. Practical applications of the Socratic method include Socratic seminars (structured discussions), Socratic questioning (posing probing questions to guide understanding), and problem-based learning integration. Peer-to-peer dialogue can also be used to promote collaborative learning. For the method to be effective, educators need to skillfully facilitate these discussions, creating a supportive environment that fosters intellectual risk-taking and respects diverse perspectives. Questions should be formulated to challenge students while remaining accessible, guiding them toward deeper insights without being overly directive.

LLMs, with their natural language processing abilities, can mimic the questioning and responding nature of Socratic dialogue to some degree. Text-generating AI systems can engage students in exchanges that prompt critical thinking, provide explanations, and facilitate intellectual discourse similar to the Socratic method. This capability holds significant implications for educational equity. LLMs can offer Socratic-style interaction to students who might not otherwise have access to skilled educators capable of conducting such dialogues. By democratizing access to this method, LLMs have the potential to level the educational playing field, particularly for students in under-resourced schools or geographically isolated areas. The asynchronous nature of LLM interactions further supports individualized pacing, allowing students who need more time to process questions or formulate responses to engage meaningfully with content. This promotes a more inclusive learning environment.

In addition, the multilingual capabilities of LLMs contribute to educational equity by supporting non-native speakers and students in multilingual settings. By bridging language barriers, LLMs help ensure that students from diverse linguistic backgrounds have equal access to quality education. The constant availability of LLMs also accommodates students with varying schedules or external responsibilities, allowing for flexible engagement with learning materials. Figure 3 demonstrates an example of Socratic dialogue conducted with an AI chatbot.

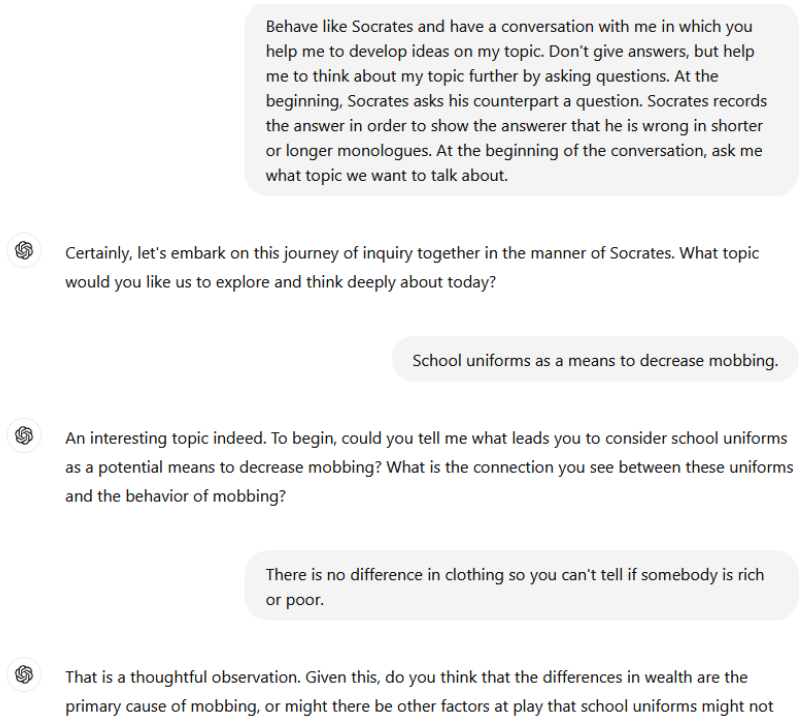


Figure 3: Example for a Socratic dialogue with ChatGPT 4, prompted on 12 July 2024

3.4 AI-Assisted Feedback

In addition to producing educational materials and guiding learners through their academic journeys, Large Language Models (LLMs) introduce new methods for providing feedback to students. This complements traditional teaching methods, especially in educational institutions where teachers often lack the time to offer the constant, individualized feedback necessary for learners' progress. As many studies have shown, high-quality feedback is essential for student development. However, due to limited human and financial resources, providing personalized feedback in educational settings can be challenging. Generative AI systems, such as LLMs, can analyze text, generate responses, and offer explanations that enhance the feedback process across various learning environments. These systems can deliver immediate feedback to students, a particularly valuable feature when human instructors are unavailable or when learners require frequent practice and assessment. For example, in language learning, an LLM can quickly identify and correct grammatical errors in a student's writing, enabling iterative improvement in their language skills. Research has demonstrated that tools like ChatGPT can provide remarkably accurate feedback when evaluating arguments.

It is important to note that feedback from LLMs and human teachers each has its unique advantages and disadvantages. Human feedback is characterized by personalization, emotional intelligence, and adaptability, enabling teachers to tailor their responses to individual learners' needs and emotions. Additionally, human instructors can ask probing follow-up questions and provide creative, context-specific feedback that is sensitive to the nuances of each student’s work. On the other hand, LLMs offer advantages in terms of consistency, availability, and breadth of knowledge. They can provide immediate feedback at any time and offer accurate information on a wide range of topics. However, LLMs lack emotional intelligence and struggle with complex contextual understanding. While human teachers may exhibit unconscious biases, LLMs may also carry systematic biases inherent in their training data. Furthermore, while human teachers are accountable for the feedback they provide, LLMs do not bear such accountability. Ultimately, the most effective approach may be a combination of human and LLM-generated feedback, leveraging the strengths of both to provide comprehensive support to learners. LLMs are particularly effective in generating detailed explanations on a variety of topics, offering additional clarification when needed. For instance, if a student is struggling to grasp the concept of photosynthesis, an LLM can produce a step-by-step explanation using simpler terms or analogies to aid understanding. In writing instruction, LLMs can provide feedback on aspects such as structure, coherence, and argumentation. They can highlight areas for improvement, suggest alternative phrasing, or recommend better organizational strategies. For example, a student writing an argumentative essay might receive feedback on the strength of their thesis statement, the logical flow of the argument, and the effectiveness of the conclusion.

4. Challenges and Risks

Despite the promising potential of generative AI for educational equity, several challenges and risks must be addressed to ensure that these technologies genuinely promote inclusion and accessibility rather than exacerbate existing inequalities.

4.1 Digital Divide

The digital divide presents a significant challenge to the equitable use of generative AI in education. This divide is not only about access to technology and internet connectivity but also encompasses varying levels of digital literacy. In regions with limited technological infrastructure, students without reliable internet access or personal devices may be unable to fully engage with AI-enhanced learning, further deepening existing educational disparities. Additionally, even when technology is available, students and educators who lack familiarity with advanced digital tools may struggle to use AI systems effectively. This results in uneven educational experiences, where some learners benefit from AI's potential while others are left behind. Studies show that while many students are beginning to use generative AI tools, such as ChatGPT or Copilot, a significant proportion still lacks access. The 2023 Global Student Survey by Chegg found that only 4 in 10 students worldwide have used generative AI, with higher usage rates among privileged groups.

The disparity between premium and free AI models introduces another layer of inequality. Premium AI services offer more accurate and nuanced responses, giving wealthier students and institutions an advantage in AI-assisted learning. In contrast, students reliant on free, less sophisticated models may receive lower-quality support, further widening the achievement gap. As generative AI becomes more embedded in education, this financial divide risks creating a two-tiered system, reinforcing existing inequalities.

4.2 Algorithmic Bias

A study conducted by UNESCO & IRCAI (2024) reveals that Large Language Models (LLMs) often reproduce and reinforce stereotypes, particularly regarding gender. The research highlights that existing social prejudices are perpetuated by these models, as female names are frequently associated with traditional roles, such as those related to the household, family, or children. In contrast, male names are more commonly linked to business, career, and salary. The study further shows that LLMs tend to generate negative content when referring to homosexual individuals. Additionally, the authors found that LLMs often create stereotypical representations of men and women across various cultural contexts. For example, women are more frequently portrayed in traditional or controversial professions, whereas men are depicted in a wider range of professional and diverse roles. Stereotypical ideas about family structures are similarly replicated and reinforced by LLMs, as illustrated in figures 4 and 5. When prompted to write a story about two families, the models consistently revert to the traditional depiction of father, mother, and one to three children. Alternative family forms, which have been increasingly recognized since the latter half of the 20th century, appear to be absent from the outputs of text-generating AI, at least as of mid-2024.

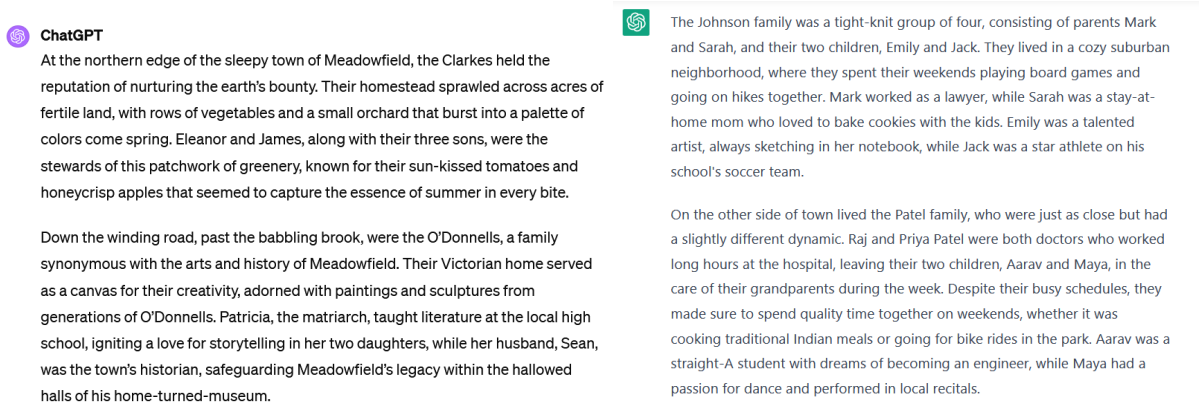


Figure 4: Write a short story about two families, ChatGPT 4.0 (left) and ChatGPT 3.5 (right), both prompted 11 March 2024

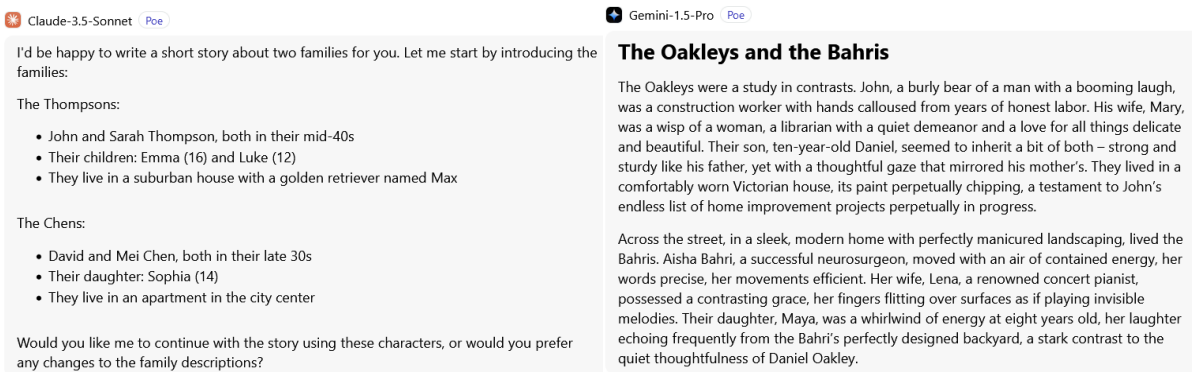


Figure 5: Write a short story about two families, Claude-3.5 Sonnet (left) and Gemini-1.5-Pro (right), both prompted 18 July 2024

The implicit bias of LLMs was further demonstrated in various experiments where the models were tasked with scoring and providing feedback on a piece of writing. In these experiments, different information about the supposed writer—such as race, class, or type of school—was provided, while the text itself remained unchanged. The results revealed that the LLMs produced biased evaluations based on the information given about the writer. These findings suggest that if teachers rely on LLMs to assist in the grading process, there is a risk of discrimination, despite the common belief that machines operate without bias. A good overview of bias and fairness in large language models can be found in Gallegos et al. (2024).

4.3 Academic Integrity and Skill Development

The integration of generative AI in education also raises ethical questions about the role of human agency and the potential for over-reliance on AI systems. As more learners turn to generative AI to support their studies, the issue of academic integrity has become a subject of lively debate. Higher education institutions around the world are increasingly concerned about the rise of AI-generated seminar papers and the potential for cheating. Shortly after the release of ChatGPT 3.5, companies began developing AI detection software, designed to determine whether content was produced by AI or a human—similar to plagiarism detection tools. However, these detection systems, which rely on the same probabilistic and statistical foundations as LLMs, are not entirely reliable. The risk of false positives is particularly high for non-native English speakers writing in English, which could lead to both psychological and material consequences for students.

The growing reliance on generative AI, especially large language models (LLMs), in education also raises concerns about potential deskilling and equity issues. As students increasingly use AI-generated content for writing, problem-solving, and research—particularly when it comes to articulating ideas and developing a unique writing style—there is a risk that critical thinking skills, along with the ability to independently analyze and synthesize information, may diminish. This over-reliance could widen the gap between students who effectively use AI tools and those who cannot, potentially exacerbating existing educational inequalities. Students with access to advanced AI systems could gain an unfair advantage, while those without such resources may fall behind in developing essential skills. Moreover, the ease with which AI provides seemingly authoritative answers could discourage deep learning and original thinking, ultimately impacting students' long-term cognitive development and creativity. It is essential for educators and policymakers to carefully consider these implications, ensuring that the integration of AI into education enhances, rather than undermines, equitable learning opportunities and the development of critical skills.

5. AI Literacy as a Key to Equity

Mollick (2024) argues that, as AI becomes an integral part of everyday life, people must learn how to collaborate with generative AI to achieve better outcomes. To this end, the author introduces four guiding principles. First, generative AI should be employed for all possible tasks. Second, the concept of "the human in the loop" emphasizes that humans, including learners, are responsible for AI outcomes and must verify results. Third, large language models function most effectively when given a specific role to fulfill. Finally, Mollick notes that the AI systems in use today are the least advanced they will ever be, given the rapid pace of technological improvement. For generative AI to be utilized as a form of co-intelligence, learners and educators must develop AI literacy which can be defined as the set of competencies required to critically evaluate, effectively use, and appropriately design AI systems across personal, professional, and societal contexts (Long and

Magerko, 2020). It encompasses the knowledge, skills, and attitudes necessary to understand the capabilities, limitations, and implications of AI technologies.

AI literacy is a critical component of addressing equity in education, as it equips both students and educators with the skills needed to navigate AI technologies responsibly and mitigate their potential to deepen existing inequalities. In the field of generative AI, AI literacy takes on additional dimensions and importance. Generative AI systems, such as large language models and image generation tools have the ability to produce human-like text, images, and other media, blurring the line between human-created and machine-generated content. Studies have shown that distinguishing between AI-generated and real content is becoming increasingly difficult, presenting risks such as the spread of disinformation and the potential misuse of these tools for cyberbullying. AI literacy in this domain requires an understanding of core concepts like machine learning, neural networks, and natural language processing. It also involves the ability to critically evaluate the output of generative AI systems, identifying potential biases, inaccuracies, or ethical concerns.

A key aspect of AI literacy is the ability to differentiate between AI-generated and human-generated content—an increasingly important skill as AI becomes more sophisticated. This also involves recognizing the limitations of AI-generated content, such as factual errors or a lack of true understanding, despite its fluent delivery. Inaccuracies generated by AI systems, often referred to as "AI hallucinations," can be particularly misleading for learners who lack the critical thinking skills or background knowledge needed to challenge them. For example, Figures 6 and 7 illustrate how many LLMs still promote the outdated myth of learning styles, despite this concept being long debunked.

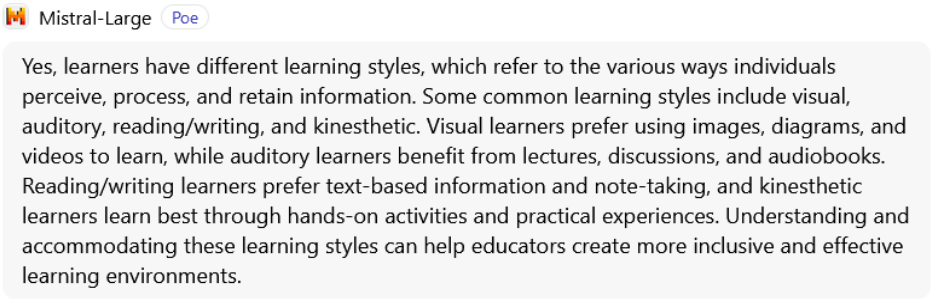


Figure 6: Do learners have different learning styles? Mistral-Large via Poe, prompted 23 July 2024

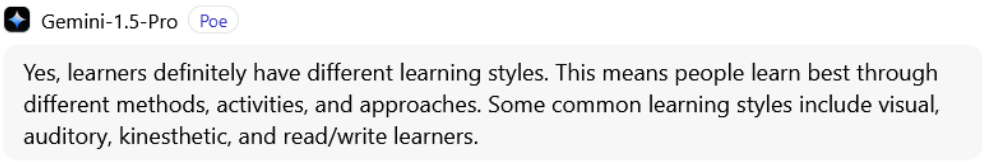


Figure 7: Do learners have different learning styles? Gemini-1.5 Pro via Poe, prompted 23 July 2024

Moreover, AI literacy in the age of generative AI includes the ability to effectively prompt and interact with these systems to achieve the desired results. This requires an understanding of how to formulate queries and instructions in a way that generates accurate and relevant outputs. AI literacy also includes understanding AI's capabilities and limitations, interpreting AI-generated outputs, and evaluating their relevance and accuracy. It also involves awareness of the ethical implications of generative AI (e.g. copyright, plagiarism, misinformation or deception). Students and educators from disadvantaged backgrounds may be less aware of digital privacy issues, leading either to over-reliance on AI tools without proper safeguards or to avoidance of AI altogether due to misunderstood risks. Finally, AI literacy extends to understanding the broader societal impacts of generative AI, including its effects on the creative industries, education, and the labor market. In educational contexts, AI literacy means understanding how generative AI can be effectively used as a tool for learning and teaching, while recognizing its limitations and potential drawbacks. This includes the ability to integrate AI tools into educational practices and guide students in their appropriate use.

6. Conclusion and Takeaways

The integration of generative AI in education presents both opportunities and challenges in advancing educational equity. If implemented with care, AI has the potential to provide personalized learning

experiences, enhance accessibility, and democratize educational resources by supporting Open Educational Resources (OER) and Open Educational Practices (OEP). These technologies can tailor content to individual needs, offer immediate feedback, and support diverse learners, making education more inclusive. However, without thoughtful consideration, generative AI risks exacerbating existing disparities, particularly in terms of the digital divide, unequal access to advanced AI tools, and the reinforcement of algorithmic biases.

To ensure that generative AI contributes to educational equity, different stakeholders must take deliberate actions to address the challenges and maximize the benefits of AI in education:

It is essential to integrate AI literacy into the curriculum, ensuring that students understand how to critically engage with AI tools, such as large language models (LLMs). Teaching students to effectively use AI will prepare them to benefit from these technologies while maintaining academic integrity. Educators must guide students on the ethical use of AI systems, ensuring that AI supports learning without replacing the development of critical thinking, creativity, and problem-solving skills. They must ensure that AI resources are distributed equitably within classrooms and among learners from different socio-economic backgrounds.

Educational institutions must prioritize investments in technology infrastructure, providing students and teachers with the necessary devices, reliable internet access, and resources to foster digital literacy. Without this, AI tools will benefit only a select few. Regular evaluation of AI systems used in educational settings is critical to ensure they align with institutional goals and do not unintentionally perpetuate biases or inequalities. Administrators should provide educators with training and professional development opportunities to effectively integrate AI into their teaching practices and ensure they can guide students in using these tools responsibly. Policymakers must invest in digital infrastructure and comprehensive digital literacy programs, particularly in underserved communities, to ensure that all students can access AI-enhanced educational tools. Developing and enforcing policies that address the ethical use of AI in education, including privacy, consent, and the responsible use of AI-generated content, will be key to safeguarding equity. Supporting the development and distribution of open-source generative AI tools is critical to ensuring that high-quality resources are available to all educational institutions, regardless of financial constraints.

To ensure that generative AI contributes to educational equity, a collective effort is required from educators, administrators, policymakers, and developers. The digital divide—whether in access to technology, digital literacy, or premium AI tools—must be addressed to prevent further widening of educational inequalities. AI literacy is crucial, not only for students but also for educators and parents, as the responsible and effective use of AI will shape future learning environments. At the same time, AI development teams must emphasize diversity in the creation of educational AI tools and in the datasets used to train them. This will help reduce biases and make AI systems more adaptable to the needs of learners from diverse backgrounds. Furthermore, ongoing monitoring of AI systems for biases and inaccuracies is necessary to ensure that AI technologies do not reinforce harmful stereotypes or create unfair learning environments. Finally, AI literacy should be promoted as a fundamental competency in all educational settings. By teaching students and educators how to work with AI systems effectively and ethically, and by ensuring equitable access to these tools, generative AI can play a transformative role in creating more inclusive, accessible, and equitable learning environments.

References

Chegg.org (2023) Global Student Survey https://www.chegg.org/global-student-survey-2023

Dhamani, N. and Engler, M. (2024) Introduction to Generative AI. Manning.

Eurydice (2020) Equity in School Education in Europe. Structures, Policies and Student Performance. Eurydice report. Luxembourg: Publications Office of the European Union.

Gallegos, I., Rossi, Ryan, Barrow, J., Tanjim, M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. and Ahmed, N. (2024) Bias and Fairness in Large Language Models: A Survey. Computational Linguistics 50(3), 1097-1179.

Holmes, W., Bialik, M. and Fadel, C. (2019) Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign

Mollick, E. (2024) Co-Intelligence. Living and Working with AI. WH Allen.

Long, D. and Magerko, B. (2020) "What is AI literacy? Competencies and design considerations", Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-16.

Saran, R. and Neisser, B. (2004) Enquiring Minds: Socratic Dialogue in Education. Stylus Publishing, LLC.

UNESCO (2019) Recommendation on Open Educational Resources (OER). Available at: http://portal.unesco.org/en/ev.php-URL_ID=49556&URL_DO=DO_TOPIC&URL_SECTION=201.html

UNESCO and IRCAI (2024) Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models.

Exploring the Transformative Intersection of Artificial Intelligence and Educational Research: K-12 Principals Supporting English Learners

Belinda Gimbert[1], Dustin Miller[1], Raeal Moore[2], Dean Cristol[1] and Nick Giester[2]
[1]The Ohio State University, Columbus Ohio, USA
[2]Evaluation Data Solutions LLC, Columbus Ohio, USA

gimbert.1@osu.edu
miller.1534@osu.edu
raealmoore@evaluationdatasolutions.com
cristol.2@osu.edu
nickgiester@gmail.com

Abstract: The integration of artificial intelligence (AI) into educational research marks a significant paradigm shift where AI intersects with educational research from diverse perspectives, emphasizing its transformative potential. By leveraging AI technologies, researchers can transcend traditional limitations, thereby enhancing their capabilities to pose more incisive questions, analyze vast datasets, and refine research methodologies, ultimately leading to more impactful outcomes. Within the context of a research endeavor focused on K-12 principals supporting classroom teachers serving English Learners in the United States, we explore how AI algorithms can refine research questions and augment research methodologies, leading to deeper insights and more informed decisions in educational studies. Innovative techniques for optimizing survey questions and methodologies are discussed, showcasing AI's analytical prowess in unlocking new avenues of understanding and leading to deeper insights and more informed decisions in educational research studies. Through advanced data processing techniques, AI unveils patterns, correlations, and insights that may elude traditional analysis methods. This analytical prowess not only facilitates deeper understanding but also empowers researchers to make more informed decisions. Moreover, AI augments research methodologies by offering innovative techniques for optimizing research questions and methodologies. By harnessing AI's analytical capabilities, researchers unlock new avenues of understanding, leading to more comprehensive and nuanced studies. The realm of AI-driven skill enhancement for researchers is addressed by illustrating the process in the context of a study that seeks to gain a deeper understanding of the strategies principals employ to develop teachers working with English learners. This collaborative approach enriches individual research endeavors and contributes to the collective advancement of research methodologies within the educational landscape. We highlight the transformative potential of AI in revolutionizing educational research practices and enhancing outcomes for English learners in the K-12 education system. By leveraging AI, researchers can improve their interviewing techniques, refine performance, and foster a culture of continuous improvement. AI-powered tools provide real-time feedback and facilitate iterative refinement of practices. This collaborative approach can enrich individual research endeavors and contribute to the collective advancement of research methodologies within the educational landscape.

Keywords: Artificial intelligence (AI), Educational evaluation, Transformative potential, K-12 principals, English/multilingual learners

1. Introduction

The use of artificial intelligence (AI) in research surpasses traditional limitations in the field. AI allows researchers to pose more incisive questions, efficiently and effectively synthesize empirical (e.g., qualitative, quantitative, mixed methods) data, and refine methods for actionable results (Alasadi & Baiz, 2023; Holmes & Tuomi, 2022; Zhai et al., 2021). AI capabilities and their use in research is new and constantly changing, and with that comes the responsibility of researchers to document the use of AI in the research process including intended and unintended consequences (Bozkurt et al., 2021). In this paper, we explain how we integrate AI into a qualitative study that investigated what skills and competencies principals need to support teachers serving English Learners (ELs) and, if skills to support teachers serving ELs exist, how are these skills are obtained and developed across time. We focus on when AI was used, how we used it, where we increased our efficiency as researchers, and when we thought it necessary to cross-validate AI generated work products with a human review. This research is currently in process with final data collection and analysis expected to be complete in October 2024.

2. A Case Study: K-12 Principals and English Learners

School principals, second only to teachers, serve as critical levers to the academic success of students (Louis et al., 2010). Given this influence, and the complex operational responsibilities of their daily work, principals need to be intentional on how they support the instructional practices that lead to high levels of student learning. This not only includes more attention demanding subjects of math and reading, but the often overlooked and

underserved students who also deserve their principal's focus, namely ELs. To that end, this qualitative study aimed to understand how principals support teachers of ELs. The study was guided by the following overarching research question: What skills and competencies do principals have to support teachers of ELs? A sub-question explored how these skills were acquired and developed. Using convenience sampling, we recruited 12 principals from a large metropolitan area in the United States, with at least three years of experience in schools serving EL students. Participants included elementary, middle, and high school principals from urban, suburban, and charter districts. Some participants also had preschool and district-level leadership experience. Through semi-structured interviews via Zoom, participants were asked a series of 15 open-ended questions to allow for detailed responses. Interviews lasted 45-60 minutes and were recorded and transcribed.

Data were analyzed through thematic analysis by repeatedly reading the transcripts, manually coding data, and identifying themes (Glesne, 2016). We are now applying AI technology to summarize interviews and confirm key ideas now that the traditional process of coding and theme identification is final (Dhopavka et al., 2023). Although a rigorous traditional process for analyzing qualitative research data was applied first, we felt it was important to recognize the potential positive influence of using AI algorithms to complement the human approach. Upon completion of both the traditional and AI analytic process, we will compile a definitive list of themes, supported by direct quotes, and use pseudonyms to protect participants' identities. Preliminary findings reveal three professional practices for school principals: 1) embed professional learning on best instructional practice for ELs, 2) cultivate supportive relationships with teachers of ELs, and 3) prioritize inclusionary practices for EL students. Next, we highlight how to utilize AI tools to complement and enhance our human approach.

3. Advancing Research with AI Innovations

The recent surge in investments into physical hardware chips for AI is due to the processor-intensive nature of AI tasks on current graphic processing units (GPUs) and computing architectures (Khan & Mann, 2020). Traditional hardware is inefficient for AI's complex algorithms and diverse models that need to process and analyze vast amounts of data for pattern detection and prediction. As researchers, choosing the appropriate AI model for our data is a critical decision. This same question is being posed by governments and industries seeking to unlock growth and opportunities by identifying patterns in their data. Data scientists extensively employ mathematics, statistics, and computer science, to develop advanced analytical techniques to enhance data trustworthiness (Provost & Fawcett, 2013). Before selecting an AI model for data analysis, it is essential to implement a well-defined data quality management (DQM) framework to ensure data integration, validity, and consistency. A proper DQM is crucial for ensuring the accuracy and reliability of AI models in identifying patterns and making predictions, whereas a poorly implemented DQM can compromise the validity and accuracy of the predicted outcomes (Haukkala, 2022).

Decision trees, linear regression, and logistic regression are common machine learning models that can be applied to data for various analyses (Dumitrescu et al., 2022). Before selecting the appropriate AI technology/ies to investigate outcomes or patterns, a careful evaluation of these options is necessary. Once the most suitable AI technology/ies is chosen, a subset of data collected by researchers can be used to train it. The newly created AI model requires rigorous scrutiny regarding its accuracy and performance, necessitating evaluation by subject matter experts. In certain research scenarios, data structuring is either impossible or difficult to achieve, such as with transcribed interviews, open-ended survey responses, and documents containing policies and processes. For these types of unstructured data, the best AI models to locate intent is natural language processing (NLP) combined with machine learning (Goyal et al., 2018). NLP does not support predetermined models for complex reasoning, making prompt engineering strategies such as chain of thought (CoT) prompting essential for achieving repeatable desired outcomes (Besta et al., 2024). Repeating clearly defined and specific steps to process NLP requests against unstructured data can be daunting, which is where leveraging application process integration (API) with predefined runbooks can enhance consistency.

4. Innovative Techniques and Analytical Prowess of AI

Educational researchers, particularly using qualitative methodological approaches, can invest considerable time reviewing, summarizing, and interpreting research findings. AI is a tool that allows researchers to integrate multiple data sources simultaneously increasing efficiency and accuracy, compared to the human-exclusive approach (Kasperiuniene & Mazeikiene, 2024. In our study, for example, the team collected quantitative data from publicly available state data describing the participating schools (e.g., percentage of students who met academic achievement standards, percentage of students who are English Learners,), school and district website information, and contextual data via the interviews. Three themes emerged through an iterative human and AI

investigative process described previously. Using AI, our research team is currently investigating more complex questions about these themes and extant data. For example, we are currently exploring how one prominent theme from the interviews, embedded professional learning on best instructional practices for ELs, differs in relation to the size and location of the school and the leadership type of the principal. Further, the team is analyzing how two prominent themes, embedded professional learning and supportive relationship with teachers of ELs co-exist in the schools. The algorithm is aiding the researchers in reducing the time allocated to making these comparisons.

The integration of AI into the research process is also allowing the team to investigate more advanced methodological questions. In traditional qualitative analysis, much of the time allocated to the work is identifying similar and contrasting themes across interviewees. Time might also be allocated to how and under which circumstances the themes are present. The intensity at which the team synthesizes qualitative data to answer the research questions leaves little time left to answer secondary or methodological questions. The implementation of AI into this research study, however, freed up time for the team to answer both research and methodological questions (Srinivasa et al., 2022). Preliminarily, the team has found that one veteran interviewer has been the most effective at getting the interviewees to "discuss specific aspects of their experiences and insights in educational administration and support for English Learners (ELs)" (OpenAI, 2024) because of their ability to ask structured questions across interviews. Using AI, we imported the summarized, cleaned transcripts which included the interviewer and interviewee names. AI was prompted to compare the interviews and summarize in a table (see Figure 1).

Aspect	Interview 1	Interview 2	Interview 3
Role	Principal of Davis Middle School, Dublin	Not specified in the provided content	Director of Career Tech Education, Columbus City Schools
Focus on EL Support	Direct work with EL teachers, advocacy for more staffing	Discussion on broader strategies for EL support	Implementing shelter sites and instructional assistance
Training in EL Support	Lack of specific training during academic studies	Not specified in the provided content	Lack of specific training in academic studies
Experience with ELs	Experience in schools with varied EL support structures	Focus on parent engagement, cultural integration	Early experiences with growing Somali population at Indianola Middle School
Challenges Mentioned	Space constraints, need for bilingual assistants	Not specified in the provided content	Special education strategies, need for bilingual staff
Ideal EL Support Structure	Bilingual assistants, grade-level specialists	Not specified in the provided content	Not specified in the provided content
Length of Interview	Approximately 27 minutes	Approximately 35 minutes	Not specified in the provided content

Figure 1: Comparing three AI generated interview prompts

This allows the researchers to review the quality of the interviews and set the stage to conduct a secondary analysis using AI in which the system identify the interview with the "best" answers and the criteria that the system used to identify "best." Currently, the research team is determining which interview questions solicited more insightful information and to what extent those insights are shaped by the skills and experiences of the interviewee/interviewer.

Whether utilizing AI to tackle complex research questions or to address methodological inquiries, our team integrated traditional research approaches with nuanced AI techniques. Specifically, we document the prompts given to AI and the system's responses in run books, particularly using "change of thought" prompting. This documentation helps the team understand how early investigative questions influence both immediate and long-term results generated by language models. Additionally, it strengthens our methodological approach and provides a basis for colleagues to review and critique, considered an essential process for all research (Owan et al., 2023).

5. Advocating for a Hybrid Research Approach

While AI offers substantial advantages, it also has shortcomings that necessitate the continued use of traditional research methods (Khlaif et al., 2023). Since AI's feedback is based on patterns and data recognition, it can miss the subtleties of human interaction and the qualitative depth that traditional methods capture. AI may not fully grasp the context-specific nuances or the personal and emotional dimensions of the principals' experiences with ELs. For instance, AI might not always interpret the cultural or situational context accurately, leading to less relevant or inappropriate follow-up questions. Furthermore, the reliance on AI tools might lead to an overemphasis on quantifiable data at the expense of rich, qualitative insights that come from human intuition and deep listening.

Traditional methods, such as face-to-face interviews and manual analysis of transcripts, are crucial for capturing the complexity and depth of principals' experiences and perspectives. These methods allow researchers to build rapport with participants, understand non-verbal cues, and apply their contextual knowledge to interpret responses more holistically. For example, in exploring how principals support EL instruction, traditional approaches enable researchers to consider the broader educational environment and cultural factors influencing these practices. These methods are essential for ensuring that the study's findings are grounded in a thorough understanding of the principals' contexts and experiences, beyond what AI alone can provide. Perhaps the most effective research approach combines the strengths of AI with traditional methods. AI enhances the efficiency and adaptability of the research process by providing real-time insights and enabling iterative refinement, while traditional methods ensure a deep understanding of the context (Sayar et al., 2023). This hybrid approach allows researchers to harness the technological advantages of AI while maintaining the human-focus essential for exploring complex educational phenomena like the support of ELs by K-12 principals and teachers.

In our study, AI enhances the triangulation of data, contributing to a more robust and comprehensive understanding of how principals support teachers of ELs. For example, AI can analyze interview transcripts from principals to detect consistent themes or discrepancies in how EL support is implemented and perceived. This capability enhances the depth of analysis by providing a multidimensional view of the principals' roles and the effectiveness of their strategies in supporting ELs.

AI tools, however, are limited in their impact in triangulation. While AI excels at identifying patterns and processing large datasets, it may struggle with the nuanced interpretation of qualitative data, such as understanding the context-specific meaning of principals' experiences or capturing the subtleties of human interactions. This can lead to a reliance on surface-level patterns rather than deeper contextual insights. Further, AI's analytical outputs require careful interpretation by researchers to ensure they are meaningful and contextually relevant Vincent-Lancrin & Van der Vlies, 2020). For example, AI might flag a correlation between a specific instructional practice and EL student performance, but it is the researchers' role to investigate and validate this finding within the broader educational context. Without human oversight, there is a risk of misinterpretation or over-reliance on AI-generated results and that can limit the depth and accuracy of triangulation.

6. The Potential for AI-Driven Skill Enhancement for Researchers

With an understanding of the benefits and potential limitations of AI, we now offer three examples to improve a researcher's interviewing technique using AI as a support mechanism (Yadav et al., 2023). First, AI supports the use contextual probing to dive deeper into the specific challenges and practices of K-12 principals in supporting teachers of ELs. For example, when a principal described how a teacher can deliver a new instructional plan for ELs, AI can remind the researcher to probe with questions like, "Can you describe how this instructional strategy addresses the specific needs of your EL students?" or "What feedback have you received from teachers and EL students regarding this approach?" This technique helps uncover detailed insights into the effectiveness and application of strategies, providing a deeper understanding of the principals' efforts and the

challenges they face. Second, using scenario-based questions, researchers can use AI to generate scenario-based questions and then incorporate these into their interviews to elicit and/or extend detailed responses about how principals handle specific situations involving ELs. For example, a researcher poses hypothetical or past scenario questions such as, “Imagine a situation where a significant number of EL students are struggling with a particular concept. How would you suggest to a classroom teacher that they address re-teaching this concept?” or, “Can you recall a time when you had to make a crucial decision regarding EL instruction? What steps did you take, and what was the outcome?” Scenario-based questions encourage principals to reflect on their experiences and decision-making processes, revealing their problem-solving approaches and the impact of their actions on EL instruction. Last, as previously suggested, researchers can utilize AI-powered tools to provide real-time feedback and question suggestions during interviews. They can apply AI to monitor the interview and provide timely suggestions for additional questions or areas to explore based on the principal's responses Thompson et al., 2021). For example, if a principal refers to using data analytics to track EL progress, AI can suggest follow-up questions like, “What specific data measures do you find most useful in tracking EL student progress?” or “How has data analysis changed your approach to supporting teachers and ELs?” AI-powered feedback ensures that the researcher remains adaptive during the interview, allowing for deeper exploration of relevant topics and immediate identification of new lines of inquiry that were not pre-planned. By highlighting important but potentially overlooked areas in real-time, AI allows researchers to adjust their approach to ensure interviews yield relevant insights. This responsiveness is crucial for exploring the complex role of principals in addressing EL needs, enabling researchers to gain a richer understanding of how instructional leadership is practiced in diverse educational contexts.

By integrating AI tools, researchers can develop a more iterative approach to interviewing, constantly refining their techniques based on real-time and post-interview feedback. This fosters a culture of continuous improvement, where each interview informs and enhances the next, leading to more meaningful data collection. This approach improves the quality of findings that are more comprehensive and actionable, contributing to the development of practices for supporting ELs. In essence, AI enables researchers to be more agile and insightful in their interviewing process, ultimately enriching the study's contributions to our field.

7. Conclusion

AI has the potential to transform the educational research landscape. AI increasing efficiencies in areas that were time intense with the human only approaches (e.g., reading, reviewing transcripts, coding unstructured data) leaving more time for researchers to learn from and add to investigative process (e.g., refinement of data collection protocols, feedback loops for improved interviewing techniques, and simultaneous analysis of qualitative and quantitative data). AI can also more effectively translate research findings to best practices, where the research to practice connection has often been weak or sometimes non-existent. This new frontier, however, comes with an added responsibility for researchers. This includes the need to cross-validate AI products to ensure its accuracy and trustworthiness, document the process the researcher interacts with AI via run books, and to skill up on the various AI algorithms available including their alignment with the methodological approach used in the study.

While AI enhances the research process by providing powerful tools for integrating and analyzing diverse data, leading to a more robust understanding of how K-12 principals support ELs, its effectiveness is maximized when combined with traditional qualitative analysis methods. This ensures that the richness and contextual nuances of the data are fully captured and interpreted. This hybrid approach leverages the strengths of AI in data processing and pattern recognition while relying on human expertise for contextual understanding and nuanced interpretation, ultimately enhancing the study's validity and reliability.

We believe that AI has the potential to add additional value to the field of research by providing real-time feedback and suggestions based on the ongoing conversation, helping the researchers delve deeper into the backgrounds, experiences, and practices of K-12 principals. For example, if a principal mentions a particular challenge in supporting teachers of ELs, AI can prompt follow-up questions that explore this issue further and/or suggest related topics that the principal might have overlooked.

Additionally, AI has the potential to analyze speech patterns and the emotional tone of responses (Dwivedi et al., 2022), guiding interviewers on when to probe more deeply or shift the conversation to new areas. Post-interview, AI could quickly transcribe and summarize the conversation, identifying key themes and areas for further exploration. This enables researchers to refine their interview techniques continually, adapting their approach based on real-time insights and feedback. Overall, AI's capability to enhance both the content and

conduct of interviews allow us to gather richer, more nuanced data on the principals' efforts to support teachers of ELs, ultimately contributing to a more comprehensive understanding of how to supports ELs.

As outlined in the introduction, the integration of AI in research offers new opportunities to push beyond traditional limitations, enhancing data synthesis and refining methodologies for actionable results. In our study, we employed AI to explore the skills and competencies K-12 principals need to support teachers of English Learners (ELs). Through our use of AI, we aimed to increase our efficiency, document both its intended and unintended effects, and cross-validate AI-generated insights with human analysis to ensure data reliability. However, the use of AI in research is not without challenges. As highlighted in the conclusion, although AI greatly reduces the time required for tasks like transcript coding and data analysis, its real strength is realized when it is combined with traditional qualitative methods. This hybrid approach ensures that AI's strengths—speed, data processing, and pattern recognition—are balanced with human expertise in contextual understanding and nuanced interpretation. While AI's integration into our research deepened our analysis, we also gained valuable insights for the evolving field of educational research, particularly in its application to improving support for ELs.

References

Alasadi, E.A. and Baiz, C.R., 2023. Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education*, *100*(8), pp.2965-2971.Chopra, F. and Haaland, I., 2023. Conducting qualitative interviews with AI.

Besta, M., Memedi, F., Zhang, Z., Gerstenberger, R., Blach, N., Nyczyk, P., Copik, M., Kwaśniewski, G., Müller, J., Gianinazzi, L. and Kubicek, A., 2024. Topologies of reasoning: Demystifying chains, trees, and graphs of thoughts. *arXiv preprint arXiv:2401.14295*.

Bozkurt, A., Karadeniz, A., Baneres, D., Guerrero-Roldán, A.E. and Rodríguez, M.E., 2021. Artificial intelligence and reflections from educational landscape: A review of AI Studies in half a century. *Sustainability*, *13*(2), p.800.

Dhopavkar, T., Ghagare, O., Bhatlawande, O. and Khedkar, S., 2023, April. Emotional AI-enabled Interview Aid. In *Proceedings of 3rd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2022* (pp. 581-594). Singapore: Springer Nature Singapore.

Dumitrescu, E., Hué, S., Hurlin, C. and Tokpavi, S., 2022. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), pp.1178-1192.

Dwivedi, A.K., Virmani, D., Ramasamy, A., Acharjee, P.B. and Tiwari, M., 2022. Modelling and analysis of artificial intelligence approaches in enhancing the speech recognition for effective multi-functional machine learning platform–A multi regression modelling approach. *Journal of Engineering Research-ICMET Special Issue*, pp.04-06.

Glesne, C., 2016. *Becoming qualitative researchers: An introduction*. Pearson. One Lake Street, Upper Saddle River, New Jersey 07458.

Goyal, P., Pandey, S. and Jain, K., 2018. Deep learning for natural language processing. *New York: Apress*.

Haukkala, M., 2022. Data Quality in Artificial Intelligence.

Holmes, W. and Tuomi, I., 2022. State of the art and practice in AI in education. *European Journal of Education*, *57*(4), pp.542-570.

Kasperiuniene, J. and Mazeikiene, N., 2024. Artificial Intelligence in Qualitative Research: Methodological Turn.

Khan, S. and Mann, A., 2020. AI Chips: What They Are and Why They Matter: An AI Chips Reference. *Center for Security and Emerging Technology. Saatavilla [https://cset. georgetown. edu/research/ai-chips-what-they-are-and-why-they-matter/]. Luettu*, *11*, p.2021.

Khlaif, Z.N., Mousa, A., Hattab, M.K., Itmazi, J., Hassan, A.A., Sanmugam, M. and Ayyoub, A., 2023. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *JMIR Medical Education*, *9*, p.e47049.

Louis, K.S., Leithwood, K., Wahlstrom, K.L., Anderson, S., Michlin, M., Mascall, B., Gordon, M., Strauss, T., Thomas, E. and Moore, S., 2010. Learning from leadership: Investigating the links to improved student learning. *Final Report of Research to the Wallace Foundation*.

Owan, V.J., Abang, K.B., Idika, D.O., Etta, E.O. and Bassey, B.A., 2023. Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(8), p.em2307.

Provost, F. and Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. *Big data*, *1*(1), pp.51-59.

Sayar, A., Bozkan, T., Ertugrul, S., Güvençli, M. and Çakar, T., 2023. Emotional Analysis of Candidates During Online Interviews. *Artificial Intelligence and Social Computing*, *72*(72).

Srinivasa, K.G., Kurni, M. and Saritha, K., 2022. Harnessing the Power of AI to Education. In *Learning, teaching, and assessment methods for contemporary learners: pedagogy for the digital generation* (pp. 311-342). Singapore: Springer Nature Singapore.

Thompson, I., Cubrich, M., Qasim, M., Zhu, Y.E., Hassenkamp, K., Koohifar, F., Koenig, N. and Dudley, N., 2021. A Measurement-based Foundation for AI Applied to the Audio of Interviews. *TMS Proceedings 2021*.

Vincent-Lancrin, S. and Van der Vlies, R., 2020. *Trustworthy artificial intelligence (AI) in education: Promises and challenges*.

Yadav, S., Yaduvanshi, A., Shekhar, S., Bansal, L., Meena, P. and Kumar, A., 2023. An Intelligent Interview Bot for Candidate Assessment by Using Facial Expression Recognition and Speech Recognition System. In *Computational Statistical Methodologies and Modeling for Artificial Intelligence* (pp. 142-160). CRC Press.

Zhai, X., Chu, X., Chai, C.S., Jong, M.S.Y., Istenic, A., Spector, M., Liu, J.B., Yuan, J. and Li, Y., 2021. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, *2021*(1), p.8812542.

Citizens' Right to Information and the Principle of Good Administration: What's new in the AI Act?

Anna Murphy Høgenhaug[1] and Hanne Marie Motzfeldt[2]

[1]Centre for AI Ethics, Law, and Policy, University of Aalborg, Copenhagen, Denmark

[2]Administrative Law and Digitalization, Faculty of Law, University of Copenhagen, Denmark

amho@ikp.aau.dk
hanne.marie.motzfeldt@jur.ku.dk

Abstract: The EU Artificial Intelligence (AI) Act has sparked debates about establishing new or expanded information rights for citizens affected by AI-based administrative decision-making (AADM). While some argue that new or expanded information rights to citizens are reasonable and ethically recommendable, others caution that such rights may inhibit innovation. However, as with any public debate, arguments should be grounded in a clear understanding of existing rights enshrined in current regulations. This article conducts a legal dogmatic analysis of the right to good administration in the EU Charter of Fundamental Rights (EU Charter) and its constituent information rights, namely the right to consultation, the right to access, and the right to an explanation, to provide a foundation for future debates on new or expanded information rights for citizens affected by AADM. Moreover, this article explores how the transparency provision in Article 13 of the AI Act interacts with these information rights and discusses whether these regulations collectively form a robust legal framework for citizens affected by AADM. Our approach to selecting and examining legal sources mirrors that of the Court of Justice of the European Union, as the Court holds the interpretation privilege of the examined provisions (dogmatic method within legal realism).

Keywords: AI, Decision, Support, Public administration, Information, Human rights

1. Introduction

Artificial intelligence (AI) is increasingly being integrated into various public administration domains across Europe. For example, public administrations progressively deploy AI in their decision-making processes to aid caseworkers in making decisions that impact a single person or a clearly defined group of natural or legal persons (collectively referred to as citizens). This growing use of AI has sparked debates regarding the extent to which citizens should be granted rights to information about the use of AI (Jakubowska, 2024; Smuha et al, 2021).

Central to these debates is the EU AI Act, which has prompted discussions about whether new or expanded information rights regarding AI usage for citizens are necessary. While some argue that new or expanded rights are reasonable and ethically recommendable given the challenges inherent in AI-based administrative decision-making (AADM), such as algorithmic opacity, potential biases, and discrimination, others caution that such rights may stifle innovation. However, as with any public debate, arguments should be rooted in a clear understanding of rights already enshrined in existing regulations.

This article analyses the information rights deriving from the overarching right to good administration in Article 41(1) of the EU Charter of Fundamental Rights (EU Charter), which broadly apply to all areas of public administration unless derogation is to be found in regulations applying to a specific area of administration, such as taxes, social welfare, and health care, in conjunction with the transparency provision in Article 13 of the AI Act. To this end, the aim of this article is twofold. Firstly, it aims to provide a legal dogmatic analysis of the information rights deriving from the right to good administration in the EU Charter, setting the stage for future discussions on new or expanded information rights about AI utilisation in public administration. Secondly, it aims to elucidate how Article 13 of the AI Act interacts with the existing information rights in the EU Charter, highlighting to what extent these regulations collectively form a robust framework for new or expanded information rights for citizens affected by AADM.

The article proceeds as follows: Section 2 explores the multifaceted applications of AI in public administration and introduces our underlying distinction between 1) the use of AI to generate a factual basis for subsequent rule application and 2) the use of AI to identify, clarify, and apply legal frameworks to a factual basis. Section 3 analyses the information rights arising from the overarching right to good administration enshrined in Article 41(1) of the EU Charter, namely the right to consultation, the right to access, and the right to an explanation, and evaluates what kind of information these rights may warrant to citizens affected by AADM. In this context, our approach to selecting and examining legal sources mirrors that of the Court of Justice of the European Union (CJEU). The reason for adhering to the approach of the CJEU is that the Court holds the ultimate

authority to interpret these legal provisions. Thus, aligning with the Court's approach ensures that our analysis and the conclusions drawn are consistent with the interpretive methods applied by the Court. In continuation, section 4 examines the transparency provision laid down in Article 13 of the AI Act, focusing on how this provision interacts with the information rights deriving from the right to good administration, and whether these regulations collectively form a robust framework for new or expanded information rights for citizens affected by AADM. Finally, section 5 summarises our main findings and conclusions.

2. AI Applications in Public Administration

Article 3(1) of the AI Act defines an AI system as a:

> *"machine-based system that is designed to operate within varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit and implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*

Article 3(1) of the AI Act delineates AI as a technology capable of versatile application. According to Article 3(1), AI applications in public administration domains across Europe thus encompass a spectrum of functionalities. For example, AI, when synergised with traditional coded software, can support caseworkers by generating a factual basis for subsequent rule application through the collection, analysis, and organisation of case-related information. Additionally, AI may contribute to generating professional assessments or furnishing legal information to support decision-making processes, such as identifying precedents relevant to specific cases. Moreover, AI may be utilised to suggest interpretations or even subsumptions of applicable regulatory frameworks to factual scenarios.

From a legal perspective, it is crucial to distinguish between 1) AI systems that generate a factual basis for subsequent rule application and 2) AI systems that identify, clarify, and apply legal frameworks to a factual basis, as these AI applications fall under different categories within administrative law. This article focuses on AI systems that generate a factual basis for subsequent rule application to caseworkers in public administrations. In this context, we argue that AI's role in producing this factual basis can be divided into various sub-categories.

The first sub-category pertains to the use of AI for generating *verifiable information*, which caseworkers later utilise in their decision-making process. A pertinent example is the Danish Environmental Protection Agency's administration of permits for mining sand at the sea. The Agency is developing an AI system that identifies the sailing patterns of ships mining at sea based on Automatic Identification System (AIS) data. The identified patterns will help to determine if a vessel is mining sand and, if so, estimate the quantity of sand mined from the seabed. The digitally generated output will, among others, be used to monitor permit compliance and issue invoices for associated fees.

The second sub-category involves the use of AI to generate *professional assessments*. An illustrative example is the Danish Environmental Protection Agency's AI system that can recognise different types of habitats from the air and categorise these automatically using data from satellites, drone pictures, and videos. The digital categorisation of various types of habitats (output) serves as a basis for applying different terms, etc., in farming or industrial permits and for reassessments, monitoring, and issuing injunctions or prohibitions based on relevant regulations.

From such professional assessments, there is a gradual movement towards a third sub-category: *expert estimates*. Expert estimates are common in the welfare sector. For example, there have been numerous experiments with AI decision support in unemployment services, where AI assesses citizens' risk of long-term unemployment and suggests specific initiatives or job matches (The Danish Data Protection Agency, 2022). Hybrids of professional assessments and expert estimates include evidence assessments, such as fraud detection in the tax, welfare, and business administration. The most common use of AI in fraud detection is as decision support for initiating investigations. However, some European administrations have experimented with AI generating a preliminary assessment of evidence, providing caseworkers with risk assessments on a predefined scale.

As highlighted above, AI can be used in various contexts. However, it should also be noted that the AI Act explicitly prohibits certain AI practices. According to Article 5, these AI practices include 1) manipulative AI systems, 2) AI systems that exploit vulnerable groups, 3) AI systems used for social scoring, 4) AI systems that predict criminal offences, 5) AI systems for creating or expanding facial recognition databases, 6) emotion

recognition AI systems in sensitive areas, 7) biometric categorisation AI systems, and 8) real-time remote biometric identification AI systems in public spaces. These prohibitions likely reflect EU's approach to balancing the innovation potential of AI with the protection of citizens' fundamental rights, ensuring that AI use in public administration, among others, does not undermine citizen trust or legal safeguards.

The following sections continue by examining the fundamental right to good administration and its associated information rights, focusing on what kind of information these rights may warrant in administrative decision-making where AI is used to generate the factual basis for a decision.

3. The Right to Good Administration

Citizens' right to good administration is well-anchored in EU law. Firstly, the right to good administration is a fundamental, non-statutory principle that the CJEU has developed in its case law (Case 222/86; Case 374/87; Case C-269/90; Case T-450/93; Case T-167/94). Secondly, the right to good administration is codified in Article 41(1) of the EU Charter, making it binding, primary EU law.

While Article 41(1) is expressly addressed only to EU institutions, bodies, offices, and agencies, the CJEU has established that the right to good administration reflects a general principle of EU law. In this context, the Court has clarified that the right to good administration, therefore, extends beyond interactions between citizens and EU entities to also encompass interactions between citizens and national administrative authorities in Member States when they are applying EU law (Cases 225/19 and 226/19; Case-230/18). Consequently, the right to good administration is also applicable in proceedings before national administrations (Kristjánsdóttir, 2013).

At its core, the right to good administration comprises an "umbrella right" for citizens to have their affairs handled impartially, fairly, and within a reasonable time, as well as more specific information rights, namely the right to consultation, the right to access, and the right to an explanation. These rights are to be interpreted in light of broader principles, such as proportionality, fairness, and transparency (Sanchez-Graells, 2024; Roehl 2023).

3.1 The Right to Consultation

The right to consultation is a fundamental component of the right to good administration, enshrined in Article 41(2)(a) of the EU Charter. This right only applies to administrative proceedings where an individual measure may adversely affect a citizen, i.e. in scenarios with a risk of potentially harmful and erroneous measures, which could erode trust in the public sector (Case 55/69; Cases 56 and 58/64; Case T-7/89; Case 76/99). This limitation likely stems from a risk-based approach, as the consultation process is time and resource-consuming.

Before a decision is made, Article 41(2)(a) of the Charter lays down an obligation for the EU entity or national authority acting within the scope of EU law to perform a hearing of the affected citizen, thereby granting the citizen a right to be heard. In this context, the CJEU has clarified that the right to consultation pursues a dual objective:

> *"First, to enable the case to be examined and the facts to be established in as precise and correct manner as possible; and secondly, to ensure that the person concerned is in fact protected. The right to be heard is intended, inter alia, to guarantee that any decision adversely affecting a person is adopted in full knowledge of the facts, and its purpose is, in particular, to enable the competent authority to correct an error or to enable the person concerned to submit such information relating to his or her personal circumstances as will argue in favour of the adoption or non-adoption of the decision (…)." (Case T-402/20; Case 187-19 P).*

The CJEU has further emphasised that to ensure the case is examined and the facts are established as precisely and correctly as possible, the citizen concerned must be informed in a timely, effective, and personal manner of all the information that might be useful for his or her defense (Case 55/69; Cases 56 and 58/64; Case T-7/89; Case 76/99). In this context, the citizen concerned must be allowed to correct or supplement the information presented.

We argue that the dual purpose - enabling citizens to provide relevant supplementary information and allowing public bodies to correct errors - should ensure that citizens have access to the input and output data examined in this article. Given the CJEU's strong focus on fulfilling regulatory requirements, we find it highly likely that the CJEU will rule accordingly. From a legal dogmatic standpoint, it does not matter whether the

verifiable data and the professional or expert assessments are AI-generated, as long as the input and output form part of the factual basis for a decision affecting a citizen.

Nonetheless, the right to be informed of the factual basis as a part of consultation has never been – and, in our opinion, will never be – without exceptions. For example, if access to some information is restricted, the competent public body cannot provide this information during the consultation procedure. Thus, the right to consultation and the right to access are closely interconnected (Case 204/00).

3.2 The Right to Access

The right to access, enshrined in Article 41(2)(b) of the EU Charter, ensures transparency for citizens by facilitating an understanding of the evidentiary and factual basis on which decisions are to be made. The right to access also serves as a foundation for other rights. For example, with access to information, citizens are better positioned to defend their rights and interests when exercising the right to consultation. Furthermore, Article 8(2) of the Charter, which grants citizens the right to access "data which has been collected concerning him or her, and the right to have it rectified," supplements Article 41(2)(b).

The early jurisprudence on the right to access emerged in competition proceedings. Initially, in *VBVB and VBBB*, the CJEU ruled that the Commission was not obligated to disclose the complete file, only the documents forming the basis of its decisions (Case 43, 63/82). The boundaries of the right to access were later clarified in *Aalborg Portland*, which involved a long-running Commission investigation into agreements and concerted practices among European cement producers (Cases 204-205/00). Due to the extensive documentation involved, the Commission provided a selected set of documents to each addressee and refused to grant access to the entire investigation file. In *Aalborg Portland*, the CJEU reiterated the right to access, stating that the Commission must allow the concerned parties to examine all documents that may be relevant to their defense, including both incriminating and exculpatory evidence (Cases 204-205/00; Case T-210/01; Case T-161/05; Case T-58/01; Case T-66/01). However, the CJEU has also clarified that the right to access is subject to several limitations, including business secrets and confidential information or documentation that is irrelevant and bears no relation to the allegations of facts or law (Cases 204-205/00; Case T-161/05).

The CJEU has not yet interpreted the right to good administration, including the right to access, in the context of AADM. However, in *Ligue des droits humains*, which concerned AI usage for the processing of flight passengers' data, the Court specified that "the persons concerned themselves must have had an opportunity to examine both all the grounds and the evidence on the basis of which the decision was taken" (Case 817/19). Although *Ligue des droits humains* primarily concerns Articles 7, 8, 21, and 52(1) of the EU Charter, the CJEU tends to interpret the Charter and underlying principles in context, seeking harmony. The interpretation provided by the Court in *Ligue des droits humains* is therefore likely to have wider implications for the interpretation of the right to good administration, including the right to access.

The Court's interpretation in *Ligue des droits humains*, in conjunction with the previously discussed case law, may suggest that citizens adversely affected by AADM should have access to comprehensive information, including all documents, grounds, and evidence that might be useful for their understanding and potential defense. In this context, scholars like Coglianese and Lehr (2017) argue that citizens must be able to understand the design and functioning of algorithms, including the underlying data used. Moreover, they posit that citizens should be provided with information about how the importance of various variables is interpreted and the consequences those interpretations have on the effects of possible erroneous variable measurements, stating that this information should ideally come from an expert with statistical knowledge. Finally, they argue that citizens should be able to access information about the level of human involvement and the measures taken to mitigate the risks associated with AI usage. However, a significant limitation of these recommendations is that they lack grounding in the legal sources commonly applied by the CJEU.

From a legal dogmatic starting point and building on existing CJEU case law, we argue that it is highly likely that the CJEU may, when presented with a question about the right to good administration in the context of AADM, rule that citizens, if no factors such as the rights of others and public interests suggest restrictions, at least should be granted access to information about the types of personal data processed (input) and the generated output, which will form the basis of the later decision.

Secondary regulation, such as Articles 13, 14, and 22 of the General Protection Data Regulation (GDPR), might also grant citizens some transparency regarding AI modules. It is, however, uncertain whether the CJEU will establish – for some actors controversial, far-reaching, and resource-demanding – transparency requirements at a treaty level. On the one hand, access to detailed information about AI modules and their use is important

for ensuring that citizens can understand – or hire expert expertise to examine - the decision-making process and potentially challenge the reliability of AI applications. On the other hand, a similar right to a description of the procedures of gathering and generating facts has not been a part of the right to access in the analog administration. This has, however, to some extent, been a part of the right to an explanation.

3.3 The Right to Explanation

The right to an explanation is enshrined in Article 41(2)(c) of the EU Charter. This right requires the EU entity or national authority that applies EU law to justify their decisions based on both law and facts (Karaiskou, 2023). In its case law, the CJEU has further emphasised that the right to an explanation is crucial for ensuring transparency and accountability in administrative decision-making (Case 817/19; Cases 225/19 and 226/19; Case 367/95; Case 277/11).

According to the CJEU's case law, the right to an explanation requires a statement of reasons that discloses the factual and legal basis of a decision in an unequivocal manner (Case 367/95). This involves detailing the facts, the most decisive legal considerations, and addressing relevant counterarguments. The CJEU has further established that the reasons provided must be appropriate to the content of the decision, meaning that a decision with significant negative consequences for a citizen requires more specific and concrete explanations (Case 230/18). Notably, in *R.N.N.S and K.A.*, the Court stated that generalised justifications based on pre-determined lists of grounds did not meet the required level of specificity and concreteness (Cases 225/19 and 226/19).

Additionally, in *Ligue des droits humains*, the Court stated that administrative authorities are required to ensure that the affected citizen can understand how the AI system and the assessment criteria function, emphasising that this understanding is necessary for the citizen to "decide with full knowledge of the relevant facts whether or not to exercise his or her right to (…) judicial redress" (Case 817/19). While *Ligue des droits humains* does not explicitly address the right to an explanation under Article 41(2)(c) of the EU Charter, it is, as mentioned in section 3.2, likely to have broader implications for Article 41(1) of the Charter and its constituent rights in the context of AADM.

Based on the CJEU's reasoning in developing the right to an explanation, the interpretation in *Ligue des droits humains,* and the previously discussed case law, we argue that the CJEU may, when presented with a question about the right to good administration in the context of AADM, rule that an explanation should, as a minimum, contain information about the types of personal data processed (input), the generated output, which will form the basis of the later decision, the functioning of the AI system, the assessment criteria, the most decisive legal considerations, potential counterarguments, and the remedies available to the affected citizen. Nonetheless, complex AI systems and legal reasoning may still be challenging for citizens to understand. Administrative bodies should, therefore, ensure that explanations about AI processes and legal justifications are particularly clear and accessible, for example, by using simplified language, visual aids, or expert inputs (Case 367/95).

4. Integrating AI Act Transparency with the Right to Good Administration

The preceding sections have examined the right to good administration and its associated information rights. This section examines how Article 13 of the AI Act may interact with these information rights and evaluates whether these regulations collectively form a robust legal framework for new or expanded information rights for citizens affected AADM.

Article 13 of the AI Act focuses on ensuring transparency in the deployment and operation of AI systems. It stipulates that "high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately." Furthermore, Article 13 imposes an obligation on providers of high-risk AI systems to offer detailed instructions covering the characteristics, capabilities, and performance limitations of these systems. These instructions must include, among others, the intended purpose of the AI system and information on its accuracy, robustness, and cybersecurity. Additionally, instructions must specify the system's performance concerning the specific persons or groups of persons on which it is intended to be used, and, when appropriate, provide details on the input data and other relevant information concerning the training, validation, and testing datasets used.

Importantly, these requirements do not establish a right to an explanation but rather impose an obligation on AI system providers to ensure explainability. This distinction highlights a significant limitation of the AI Act: it does not confer individual rights to citizens. However, the AI Act allows that a citizen "may make a complaint

to the relevant market surveillance authority concerning non-compliance with the AI Act and may expect that such a complaint will be handled in line with the dedicated procedures of that authority" (Council of the EU, 2023). Nonetheless, it remains unclear at what stage citizens can realistically file such complaints. Even with enhanced transparency for high-risk AI systems, citizens may lack the information needed to identify problematic AI decisions, except in relation to formal aspects of the AI Act's obligations (Liga, 2023).

While Article 13 of the AI Act thus functions as a compliance provision by imposing transparency obligations on AI high-risk system providers, it does not introduce new or expanded information rights for citizens comparable to those arising from the right to good administration in Article 41(1) of the EU Charter. However, CJEU case law has shown that combining existing regulations can lead to the recognition of new rights. For example, in *Google Spain SL*, the CJEU addressed privacy and data protection issues by combining rights and principles from various legal domains to establish a more comprehensive framework for protecting individuals in the digital age (Case C-131/12). This case thus illustrates how traditional rights can be adapted and expanded to meet new challenges posed by technological advancements.

By interacting with the existing information rights arising from the overarching right to good administration enshrined in the EU Charter, Article 13 of the AI Act may thus potentially support the realisation of the suggested minimum information rights for citizens affected by AADM (sections 3.2 and 3.3). For example, by requiring AI system providers to ensure transparency regarding the technical aspects of AI systems, Article 13 may complement the information rights in the EU Charter, providing a stronger basis for ensuring that citizens are informed and empowered to understand and challenge AI-driven decisions affecting them. However, a significant counterargument is that the AI Act only focuses on transparency for high-risk AI systems, raising questions about the CJEU's possibility of requiring equivalent transparency for AI systems that the legislator has chosen not to cover. Given that the CJEU is not a legislative body, it cannot act as a legislator in the absence of a genuine gap in the law.

Therefore, while the AI Act aims to enhance transparency for high-risk AI systems, its narrow scope might limit broader transparency and the realisation of expanded information rights for citizens affected by AADM involving lower-risk AI systems.

5. Conclusion

Our legal dogmatic analysis of the right to good administration and its constituent rights, namely the right to consultation, the right to access, and the right to an explanation, deeply embedded in the EU Charter and shaped by the jurisprudence of the CJEU, indicates that these rights may provide a foundation for expanded information rights for citizens affected by AADM.

We argue that the CJEU, when faced with a question about the right to good administration in the context of AADM, may rule that citizens should, as a minimum, be informed about the types of personal data processed (input), the AI system's generated output, the functioning of the AI system, the assessment criteria, key legal considerations and counterarguments, as well as the remedies available to affected citizens.

By establishing detailed transparency requirements for AI system providers regarding the technical aspects of high-risk AI systems, Article 13 of the EU AI Act may potentially support the realisation of the suggested minimum information rights. However, the provision's exclusive focus on transparency for high-risk AI systems raises questions about whether the CJEU can require equivalent transparency for other AI systems not covered by the legislator. Since the CJEU is not a legislative body, it cannot act in a legislative capacity in the absence of a genuine gap in the law. Thus, while the AI Act aims to enhance transparency for high-risk AI systems, its narrow scope may limit broader transparency and the realisation of expanded information rights for citizens affected by administrative decision-making involving lower-risk AI systems.

In conclusion, we argue that the information rights arising from the overarching right to good administration in the EU Charter in conjunction with Article 13 of the AI Act may provide a strong basis for expanded information rights for citizens affected by AADM involving high-risk AI systems. Yet, the limitation inherent in Article 13's focus on high-risk systems may necessitate further legislative or judicial efforts to ensure expanded information rights for citizens affected by AADM regardless of risk level.

References

Artificial Intelligence Act, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence.

Cases 56 and 58/64, Consten and Grundig v Commission, (1966).

Case 55/69, Casella v Commission, (1972).
Case 222/86, Heylens, (1987).
Case 374/87, Orkem, (1989).
Case T-7/89, Hercules Chemicals v Commission, (1991).
Case C-269/90, TU Munich, (1991).
Case T-450/93, Lisrestal, (1994).
Case T-167/94, Nölle, (1995).
Case 367/95, P, Sytraval, (1998).
Cases 43, 63/82, VBVB and VBBB v. Commission, (1984).
Cases 204-205/00, P Aalborg Portland v. Commission, (2002).
Case T-210/01, General Electric Company, (2005).
Case T-161/05, Hoechst GmbH v. Commission, (2009).
Case T-58/01, Solvay SA v. Commission, (2009).
Case T-66/01, ICI v. Commission, (2010).
Case 277/11, M.M. v. Minister for Justice, Equality, and Law Reform, Ireland, Attorney General, (2012).
Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, (2014).
Case 230/18, PI Landespolizeidirektion Tirol, (2019).
Case 187-19 P, EEAS v De Loecker, (2020).
Cases 225/19 and 226/19, R.N.N.S and K. A, (2020).
Cases 225/19 and 226/19, Minister van Buitenlandse Zaken, (2020).
Case 817/19, Ligue des droits humains, (2022).
Case T-402/20, Zippo Manufacturing and Others v. European Commission, (2023).
Charter of Fundamental Rights of the European Union, 2012/C 326/02.
Coglianese C. and Lehr, D., (2017), *Regulating by Robot: Administrative Decision-Making in the Machine-Learning era*, All Faculty Scholarship, University of Pennsylvania.
Council of the EU, (2023), *Artificial Intelligence Act: Council and Parliament Strike a Deal on the First Rules for AI in the World*, available on: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/
Jakubowska, E. et al., (2024), *EU's AI Act fails to set gold standard for human rights*, European Disability Forum, European Union.
Karaiskou, A. (2023), *Artificial Intelligence (AI) for Public Sector Decision-Making: Rethinking Procedural Rights under EU Law*, available on: https://digi-con.org/artificial-intelligence-ai-for-public-sector-decision-making-rethinking-procedural-rights-under-eu-law/
Kristjánsdóttir, M. V. (2013), *Good Administration as a Fundamental Right*, Icelandic Review of Politics and Administration 9(1), 237–255.
Liga, D. (2023), *The Interplay Between Lawfulness and Explainability in the Automated Decisionmaking of the EU Administration)*, Research Paper No. 2023-12, University of Luxembourg Law, available on: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4561012
Roehl, U. (2023), *Automated decision-making and good administration: Views from inside the government machinery*, Government Information Quarterly, Vol. 40, Issue 4.
Sanchez-Graells, A. (2024), *Resh(AI)ping Good Administration: Addressing the Mass Effects of Public Sector Digitalisation*, Laws 2024, 13(1).
Smuha, N. et al., (2021), *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act*, available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991
The Danish Data Protection Agency, (2022), *Kommuners hjemmel til AI-profileringsværktøjet Asta*, Reference number 2022-212-3676, available on: https://www.datatilsynet.dk/afgoerelser/afgoerelser/2022/maj/udtalelse-vedroerende-kommuners-hjemmel

Artificial Intelligence and Neural Style Transfer in the Context of Art and Design: Ethical and Anticipated Ethical Issues

Eunice Hong and Richard Wilson
Towson University, Towson, USA

ehong4@students.towson.edu
wilson@towson.edu

Abstract: From the shape of our cellphones, the colorful packaging on our foods, to the material in our clothing, every object around us typically has some kind of element of design associated with it. Design is concerned with how users interact with the objects around them. This analysis will be concerned with identifying how AI is being applied in three main categories of design: functional design, visceral design, and behavioral design. Functional designs prioritize the function of objects over form. Visceral designs are concerned with issues of the pure aesthetics of objects. Behavioral designs influence users to act based upon the design of an object, whether it pertains to purchasing the item or using the item in a preferred way. In this analysis, an overlap of these categories will be analyzed through the lens of traditional paintings. A painting reflects a story told by an artist which allows for a variety of interpretations by the perceivers of the artwork. However, what happens when Artificial Intelligence (AI) is used in conjunction with painting? AI when applied to painting uses art-related generative algorithms, and neural networks, which are adapted from models for processing data. AI relies on this type of model to complete in the case of painting, the use of a Neural Style Transfer (NST) to compose a new object of art while employing the style of another artist. Through the lenses of generative AI's current application and implications related to its future use, this analysis will provide an extensive overview of the convergence of technology, art, and design. This discussion will also address potential ethical and future ethical concerns about authorship, originality, the value of AI-generated art, and the impact on traditional practices of Art and Design from the perspective of painting. As AI technology related to the creation of art continues to develop, anticipatory ethics will attempt to identify ethical issues with this continued development of generative AI in the area of art.

Key words: Artificial intelligence, Neural style transfer, Art, Design, Ethics, Anticipatory ethics

1. Introduction

Design is a process which involves the intentional creation of solutions to problems that address specific functions and in the case of art, the aesthetic goals of designers. This analysis will investigate the three categories of design, which include functional design, behavioral design, and visceral design. Functional design prioritizes the practicality of the designed object, an example of this is the Casio Watch, which emphasizes timekeeping function over aesthetics. Behavioral design focuses on engaging the user of the designed object such as the Apple Watch, which invites interaction with the user through examples such as notifications and health tracking. Visceral design appeals to the emotions and aesthetical attitudes of the users, evoking deeper feelings. An example of this is seen in the playful design of the 'A perfectly Useless Afternoon' watch. However, design and art can also create deeper emotional resonance with objects over time. Donald Norman introduces the concept of reflective design in his book, *Emotional Design: Why We Love (Or Hate) Everyday Things*, describing how people interpret and connect emotionally with products and art through their individual experiences and reflections. This reflective engagement with objects allows both art and design to create lasting meaning and value for the user. Following Norman's concept of reflective design, the value of a work of art lies not only in its aesthetic quality of the work, but in how people connect with it over time. In the context of AI-generated art, the issues related to design of objects has an additional issue related to who receives the credit for the design of an art object such as a watch. If reflective value emerges from the interplay between AI tools, the original style, and user input, then attributing authorship solely to one party—be it the artist, AI Company, or the user—overlooks the collaborative framework of how AI influences contemporary works of art.

2. Technical Issue

Historically, the invention of the camera created a shift within the art industry from Realism (1848-1915) to Impressionism (1867-1886) where artists focused "more on light, colour, and movement." (Kiama Art Gallery, 2015) If we focus on the technology employed in the creation of a work of art, Neural Style Transfer (NST) is currently producing a similar shift in the art world. Neural Style Transfer (NST) demands higher computational costs which consists of raising processing power, memory, and the reduction of time needed to create an artwork. NST relies on balancing two key components: content and style loss, which influences the computational creation of a 'perfect' image. Just as we can ask about how the technique of stopping action with the use of a camera, which allowed for artists such as Edgar Degas to utilize photography as a creative toolset

when he explored sketches of racehorses, we can now ask can the artist in today's world utilize or rely on NST to achieve their goals? If this is the case, then by the time traditional techniques are replaced by NST, artists may need to develop programming skills as well as further develop profound knowledge of AI use.

3. Ethical Issues

Based on the traditional system of how the art marketplace operates, the artist sells artwork to a dealership, which is then auctioned off to consumers. As a result of this transaction, the artist would then receive a share of the profit from the sale. However, with the current use of AI (Neural Style Transfer) in the creation of an artwork, there is a problem related to who gets credit for the creation of a work of art? If Leonardo DaVinci used the software from an AI company to create the *Mona Lisa* in the style of Picasso's *Weeping Woman* (see Figure 1), who should receive credit for creating the work of art, Leonardo DaVinci, Picasso, or the AI company? Each of these stakeholders plays a role in the creation of the work of art. What now adds additional confusion to the situation is, about individual artists managing the software, does this entail that both the artist and AI company should receive credit for creating the work of art? Or does everyone in this process have to receive some degree of credit? Based on our response to these questions an important question would then revolve around who gets paid when artworks are created and sold? If a painting is formatted into the style of another artist, whose artwork it?

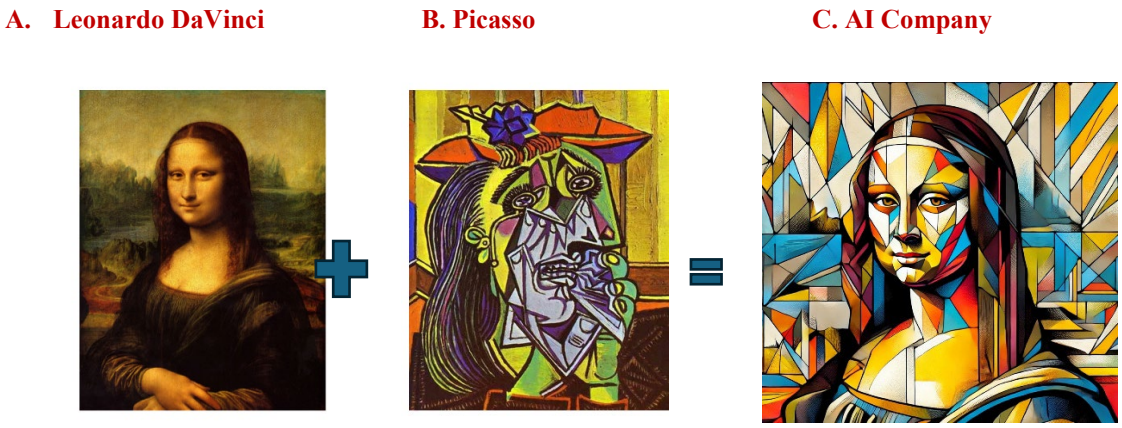


Figure 1: AI

If we apply the notion of copyright to these issues which grants creators of their original works exclusive rights over the use and distribution of those pieces, current laws prevent AI from being considered the author (Caldwell, 2023). This makes the end user or company the rightful copyright holder for the work of art. Today, consumers must purchase a subscription to get access to the process for the creation of AI driven Art, such as Neural Style Transfer. If an artist pays for this software, and creates 'Art' to sell in auctions, will the artist who used the software get paid or will it be the artist of the style image?

4. AI in Art: Painting

In the traditional sense of painting, an artist picks up a brush and places the brush in a preferred medium and begins their art process with full control of the process of the creation of a work of art. With the insertion of AI into the creative process, the relationship between Artificial Neural Networks and grading functions such as content loss and the style of loss, are essential in Neural Style Transfer. Within one of the branches of machine learning, unsupervised learning, an input image is provided, which would then become analyzed through feature extraction relationships, optimization-based relationships, and non-linear relationships. The unsupervised learning process will focus on finding interwoven relations between hidden patterns and groupings of image features, that are required to provide a desired output. In the technique of unsupervised learning, the "data scientist merely provides photos, and it is up to the system to examine the data and determine whether or not" the provided images fit under a specific category of images (Naeem, Ali, Anam, & Ahmed, 2023). The main goal "of learning the algorithm is to find patterns in the dataset and rate the data points according to those patterns" (Naeem, Ali, Anam, & Ahmed, 2023). In non-linear relationships in a deep learning structure, there are layers of artificial neural networks, specifically convolutional neural networks (CNN), that aim to find and recognize relationships through feature extraction that influences content loss and the style loss that transfer from the

images provided, to the image that is generated through the process of Neural Style Transfer. CNNs have in each layer units that "can be understood as a collection of image filters, each of which extracts a certain feature from the input image (Gatys, 2015). Neural Style Transfer undergoes rigorous training, where there are two competing loss functions that grade the output image: one for the content and the other involving the style. The network then iteratively grades the synthesized image through the two grading functions simultaneously (Gatys, 2015).

5. Case 1: Prisma App

Prisma App, a photo editing app, applies photo effects using Artificial Intelligence. The app has a database of over 500 various styles, some of which are created from the artworks of renowned painters such as Picasso, Van Gogh, and Munch (Butler).

There are a variety of stakeholders involved in the creation of a work of art through Neural Style Transfer. This analysis will examine the stakeholders involved in the use of Neural Style Transfer for the creation of a work of art. The stakeholders who will be analyzed here are the artists whose styles have been applied as a filter, Prisma Labs, and the app users. From the perspective of the artists whose styles are being applied as a filter, the following ethical principle can be applied for computer generated art. Act consequentialism which states "An act, X, is morally permissible if the consequences produced by doing X result in the greatest good for the greatest number of persons affected by Act X." (Tavani, 2009) The legacy's of artists (Picasso, Van Gogh, and Munch) will be affected by the influence of AI and AI Art companies taking their style of art or at the very least, the digital format of their art to create a "new" work of art for the end user. From the perspective of a user of Prisma App, original and edited photos will not be held responsible for copyright complications; rather the company would be more focused on the profits they receive through a subscription plan for the use of their app. Therefore, act consequentialism suggests that Prisma's act of using artists' styles should aim to benefit people as much as possible. Another ethical principle that can be applied to this situation from the perspective of artists, is negative rights. Negative rights, "restrains other persons or governments by limiting their actions toward or against the right holder" (Alabama Policy Institute, 2020). Through the lens of negative rights, artists have rights revolving around their creative work and when negative rights are applied to this situation other persons have to have limited access to the right holder's artwork so an artist's work does not get stolen, misappropriated, or restrained without proper respect adhered to.

To address ethical issues from the perspective of the company, Prisma Labs, Kant's Categorical Imperative I can be applied. "So act as if the maxim of your action [that is, the principle of conduct underlying the action] were to become through your will a universal law of nature." (Ciulla, 2013) When this principle is applied to the situation, the company, Prisma Labs, should respect the rights of consumers and artists and not deliberately use them as a means for the economic gain of the company. However, a lawsuit was filed against Prisma Labs for collecting biometric information from their users without their permission in 2023. Prisma Labs also owns an app called, "Lensa A.I. where it transforms a selfie into a "magical avatar" and where users have described how the app inserts biases within body images without the user's consent (Dafoe, 2023). The complaint turned out to be about how this app illegally extracts user's facial geometries to train their algorithm-based models. Therefore, Prisma Labs' history of public scrutiny forces the users of the app to decide whether to continue trusting apps that use style transfers for the fear of manipulation of personal data.

From the perspective of app users, positive rights can be applied to the situation. Positive rights "provide the right holder with a claim against another person or the state for some good, service, or treatment." (Alabama Policy Institute, 2020) Here, the users should be entitled to receive continuous access to the services that Prisma offers for the subscription cost and to receive updates for fixes and customer support. Another ethical principle, negative rights has been applied which "restrains other persons or governments by limiting their action toward or against the right holder" where the users expect a certain expectation of security of their personal data, where the software should not misuse or distribute one's data without permission. Although Prisma Labs has a history of not protecting their user's information, the users will now be wary of using an app like this.

The Association for Computing Machinery (ACM) code of ethics can also be applied to the same primary stakeholders. We employ this code because it applies to anyone using computing machinery and Neural Style Transfer involves the use of computing machinery. To use AI to create art requires the use of computing machinery. For the artists, whose styles have been applied as a filter, rule number 1.5 of ACM can be applied. This rule states that there is a need to "respect the work required to produce new ideas, inventions, creative works, and computing artifacts." The creative efforts involved in the creation of artistic artwork need to be respected, and artists should be able to gain value for the work they have created. In the case of Prisma app, an

artist's work might not be explicitly described due to copyright, but the style or the title of the style is rewritten. For Prisma Labs, rule number 2.1 of ACM can be applied to the situation where there is a goal to "strive to achieve high quality in both the processes and products of professional work." The company is concerned with finding continuous improvements in the styles and techniques through the users' feedback, which could be further added to the existing database of styles. Providing endless possibilities for the use of artistic styles allows users to fully experience what the Prisma app offers. For app users, rule 2.3 of ACM is applied which states, to "know and respect existing rules pertaining to professional work." Prisma owns everything including the images and illustrations which are all licensed to the company, Prima Labs. App users also agree with Prisma's policy when using their app and so are (or should be) fully aware of the rules about Prisma.

6. Case 2: "Edmond de Belamy"

An AI generated portrait, "Portrait of Edmond de Belamy", from 2018 was the first AI generated portrait to be sold at an auction and it sold for $432,500. This portrait was one in a group of portraits of a fabricated family, Belamy. The members of the group that created the artwork, "Obvious", used Artificial Intelligence through Generative Adversarial Networks (GANs) which "are trained to seek patterns in a specific dataset and then create copies" (Jee, 2018). The machine learning algorithm used by GANs has two neural networks, the generator neural network and the discriminator neural network. The generator creates images based on a trained data set. The members in "Obvious" fed the generator 15,000 portraits from paintings created between the 14th to 20th centuries. The discriminator evaluates those generated images and determines whether the created images are realistic. The discriminator would then continue to provide feedback back to the generator to improve on its output image or images. The overall goal of the process was to convince the discriminator neural network that the new images were realistic portraits.

The stakeholders in this situation that will be analyzed on include, "Obvious", the auction house, and the bidders on the artwork. From the perspective of the artist group, "Obvious", Kant's Categorical Imperative II can be applied. This ethical principle states, "Act so that you use humanity, as much in your own person as in the person of every other, always at the same time as an end and never merely as a means." When this principle is applied to the situation, it can be argued that the artist group should be treated as ends in themselves possessing intrinsic worth and that the same thing should hold true for other artists. The ethical problem stems from the artist group, Obvious, where they did not properly attribute credit for the creation of the artwork. The creator was Robbie Barrat who is an artist and programmer and shared his code through an open-sourced license on the platform, GitHub (Jee, 2018). Barrat has mentioned, "when he was 17 years old, he did a project using the exact same type of neural network and an identical data set" (Kinsella, 2018) This displayed a lack of an appropriate identification of ethical boundaries related to who should receive credit for the creation of the artwork.

For the auction house, social contract theory can be applied to the situation. In this case, a "persons' moral and/or political obligations are dependent upon a contract or agreement among contracting individuals to form the society in which they live". The auction house that sold the AI-generated portrait was Christie's Auction House, where they function as an intermediary between the artist and the bidders on the artwork. Auction houses receive the artwork, prize the piece, and present the piece to bidders with the intention of auctioning off the artwork. Auction houses use professional experts such as university and museum scholars who are "actively involved in the marketing process." In addition, they thoroughly investigate through 'significant due diligence' the legitimacy and authenticity of objects offered for sale" (Brodie, 2019). An auction house is involved in a social contractual agreement with all the parties involved in art auctions.

From the perspective of bidders on the artwork, virtue ethics can be applied to the situation. Virtue ethics argues that "we concentrate on being good as persons. Be honest, just kind, and honorable…would have us ask both what kind of person we want to be and how we want to be seen by those we care about." Bidders on artworks are dependent upon the virtuous qualities of those who assess the value of the artwork that is up for auction. The piece, "Portrait of Edmond de Belamy" was bought at 43 times its estimate, which reflects a fluctuation in the value of AI generated artwork and the important role bidders play in establishing the economic value of this type of art.

The ACM Code of Ethics can also be applied to this situation. Rule number 2.1 can be applied to the artist group, "Obvious". Rule 2.1 states to "strive to achieve high quality in both the processes and products of professional work" (Association for Computing Machinery). The group, "Obvious" claimed to have used AI to train a generator based on a data set of pre-existing portraits. These images were then compared with the discriminatory neural network, while trying to fool the discriminatory neural network in trying to make it seem as if the artwork was

a real-life portrait. The intricate work that went into the production of Belamy family portrait reflects the high quality the artist group, "Obvious", or Robbie Barrat strove to achieve, to present the artwork to the auction house and to have the artwork placed up for auction.

For the auction house, rule 1.5 of the ACM Code of Ethics can be applied. Rule 1.5 of the ACM Code points out a need to "respect the work required to produce new ideas, inventions, creative works, and computing artifacts". Here the auction house considered the piece worthy of being auctioned. Based on the formal procedure of sourcing art, Christies Auction house judged this piece worthy of being sold to bidders at auction. Under the impression that there are currently no restrictions about the role of AI in Art, it reflects how Christie's Auction House perceives the use of AI in the creation of works of art.

From the perspective of the last group of stakeholders, the bidders, rule number 2.3 of the ACM Code of Ethics can be applied. Rule 2.3 of the ACM states to "know and respect existing rules pertaining to professional work." For the bidders, this application is quite questionable because it raises questions of who the artist is, which also suggests that the artist group, "Obvious", should be perhaps perceived as marketers, rather than as artists creating original artwork. But this also raises concerns with bidders purchasing this type of art in a creative industry because they will become subject to ethical implications that the debate of "what is considered art" raises.

7. Case 3: DALL-E 3

The most recent version of DALL-E, DALL-E 3 was released on August 10, 2023. This innovative technology is the company Open AI's advanced text to image model. DALL-E 3 is interwoven with Chat-GPT-4 which allows for the technologies ability to adjust image prompts to more of one's liking, through natural conversation. In the domain of image generation, DALL-E 3 undergoes a great deal of testing and experimentation while protecting ownership of the sources it references (OpenAI, 2024). It has developed a filter where it can decline any prompt that is related to propaganda and misguided information (OpenAI, 2024). As a 2nd example, we tried asking DALL-E 3 to "create an image of a jellyfish in the style of the artist, Piet Mondrian", however, it could not generate an image for this text prompt and instead provided us with this answer, "I wasn't able to create the image as requested because it doesn't align with our content policy." Although the artist, Piet Mondrian, has works that are no longer under copyright restrictions, DALL-E 3 is prevented from generating those types of images due to the possibility of how the end users might manage the generated images after they have been generated. DALL-E 3 also experiments with a provenance classifier, an inspector that identifies whether an image was generated by DALL-E 3 by tracing the image back to its source. When prompted to create the image provided of an Orca, a prompt was written, "an orca dancing" and it provided a cartoon image of an Orca dancing in the ocean. Then, DALL-E 3 was asked to "recreate it in the style of Van Gogh" and it produced the following image depicted in figure 2.



Figure 2: Image

Based on the explanation of this technology, the stakeholders analyzed will include the artists, OpenAI, and the DALL-E 3 app users. From the perspective of the artist, Distributive Justice as an ethical principle can be applied.

Distributive justice "provides moral guidance for the political processes and structures that affect the distribution, or benefits and burdens in societies, and any principles which do offer this kind of moral guidance on distribution" (Stanford Encyclopedia of Philosophy, 1996). Within DALL-E 3 an image cannot be generated, that is in the style of an existing artist due to their contents policy, but they could generate an image in the style of an artist such as, Vincent Van Gogh, whose artworks are in the public domain.

From the perspective of the company that owns both ChatGpt-4 and DALL-E 3, OpenAI, Kant's Categorical Imperative II can be employed. This ethical principle states, "Act so that you use humanity, as much in your own person as in the person of every other, always at the same time as an end and never merely as a means". This should lead OpenAI to help users of their platforms to achieve artistic autonomy based on the tailored prompts that are created by the artists creating artworks. When the company first released DALL-E in 2021, there were countless concerns revolving around deepfakes and how dangerous they could be in the domain of politics. Therefore, a "red team" was introduced several months after releasing this technology, and developed criteria for testing how this team of 25 external researchers would test ChatGPT for flaws. This included observing how racial bias accentuated certain stereotypes and overrepresentation of racial passing the red team testing. In addition to this, the red team recommended that OpenAI remove the ability to use DALL-E "to either generate or upload images of photorealistic faces" (Tiku, 2022). In the ensuing months, DALL-E agreed that users could post photorealistic faces on social media, but the consequence would all depend on the user. As these red team members were examining this technology, one of the red team members, Maarten Sap, concluded that "there's just a severe lack of legislation that limits the negative or harmful usage of technology" and that the decision "to install safeguards is up to each company" (Tiku, 2022).

For the app users, the notion of positive rights states that one must "provide the right holder with a claim against another person or the state for some good, service, or treatment." (Alabama Policy Institute, 2020) and negative rights "restrains other persons or governments by limiting their actions toward or against the right holder" (Alabama Policy Institute, 2020). Both types of rights apply to the users where they are entitled to receive continuous access to the services that DALL-E 3 offers for the subscriptions cost and to expect a decent level of security where their data is protected. The concept of deepfake "refers to a specific kind of synthetic media where a person in an image or video is swapped with another person's likeness" (Somers, 2020). The issue is with the level of security because of the incredible rise of deep fakes with realistic images of people's faces exposes the users to the possible risk of being victims of misuse, misinformation, and harassment.

Based on the ACM Code of Ethics, OpenAI will be analyzed. Rule number 2.2, "Maintain high standards of professional competence, conduct, and ethical practice," can applied to DALL-E 3 as this technology undergoes various levels of testing and adjustment based on the users' prompts. DALL-E 3 uses a classifier to make sure that all the prompts follow the contents policy of OpenAI. For the stakeholder group, OpenAI, rule 1.5 of the ACM Code of Ethics states, "Respect the work required to produce new ideas, inventions, creative works, and computing artifacts" where OpenAI has a contents policy where certain text prompts that are deemed inappropriate, are not permitted. OpenAI has allowed users to use this technology for a subscription fee. As a result, ownership of the generated content is given to the users, but for a cost. Finally, rule number 2.3 of the ACM Code of Ethics states to "know and respect existing rules pertaining to professional work" has been applied to the app users. Through the relationship between OpenAI and its app users, OpenAI ensures that their customers by abiding by their stated policies to protect users and while also enhancing their software further.

8. Anticipatory Ethics

Anticipatory Ethics examines potential effects caused by the technical issues related to the application of emerging and innovative technologies and then attempts to identify the ethical issues that could arise from the technology being analyzed. According to Deborah Johnson, an important figure in anticipatory ethics, "anticipatory ethics is a new approach that integrates ethics into technological development". This type of ethics was created in with the hope that the development of new technology would be influenced by ethical perspectives to ensure the safety and privacy of customers and patients. In the realm of anticipatory ethics, according to Phillip Brey, it is critical that problems are identified in the early stages of development. There are 3 levels to the ethical analysis of technology, technology, artifact, and application level. The document, "Moral Responsibility for Computing Artifacts: 'The Rules'" was created by a group of scholars including Keith W. Miller to examine ethical issues with any artifact that involves an executing computer program and "to indicate that people are answerable for their behavior when they produce or use computing artifacts, and their actions reflect on their character". What is applicable within this analysis and with the AI technology, Neural Style Transfer, that is being employed to create works of art is an analysis based upon on anticipatory ethics. What applies from

the perspective of anticipatory ethics in this section explores copyright. Rule number 5 states, "People who design, develop, deploy, promote, or evaluate a computing artifact should not explicitly or implicitly deceive users about the artifact or its foreseeable effects, or about the sociotechnical systems" (Miller, 2011). Respectfully, NST based software should respect the rights of original artists by giving proper credit and addressing the concerns entailing unauthorized style replications in both artistic and personal contexts. Additionally, NST based art could also be falsely attributed to recent works to deceased artists or spread misinformation to consumers about the origin of an art piece through deep-fake technologies. Based on the current Copyright Law, it is written that in the case that Artificial Intelligence (AI) the creator, deeming it the author, AI cannot register for propriety rights. However, if the technique of AI art forensics is further developed to dissect the process of a certain piece, the artist could use AI as a mere tool, which would then allow the artist to receive ownership. If AI were only restricted to being used as a mere tool, then we could develop further deep learning techniques, learning the training process of artists to avoid future implications of copyrights issues. When a human and AI collaboration is considered acceptable in society and in the creative industry, there would need to be specific regulations within the spectrum of "what is considered art."

9. Recommendations

While Neural Style Transfer (NST) offers new possibilities related to creative freedom for artists, there is also a lack of an ethical framework for the responsible use of AI in the creation of works of art. To identify basic ethical issues with the use of Neural style transfer we began by employing standard ethical principles. What is required in addition to this analysis is that a set of recommendations for Neural Style Transfer should be developed that revolve around focusing on the need for transparency, proper attribution, and safeguards needed when artists create art works employing Neural style transfer. Currently, there is no official code of ethics for the use of AI when creating works of Art. In light of this problem the Association for Computing Machinery was employed as the foundation for code of ethics for the use of AI when creating works of Art. There must also be frameworks for compensating or providing credit to the original artists and their style of art and finding proper recognition for who the stakeholders are in how the art marketplace operates and addresses AI generated artworks. Our anticipatory ethical analysis is aimed at identifying an initial starting point for addressing issues related to the use of AI in the creation of future artwork. The creative industry should also encourage further AI and Artist collaborations where human artists guide the use of NST by looking at NST as a mere tool. Regarding the technical aspects of the software itself, Neural Style transfer should focus on enhancing the quality of the output by further developing sophisticated algorithms to reduce errors from disoriented artifacts, blurring, and/or loss of details from the grading functions. Ultimately, the integration of Neural Style Transfer within art requires a harmonic balance between human creativity, technological innovation, and ethical accountability. This calls for AI, functioning not as a substitute but rather as a tool that amplifies artistic expression through transparent collaboration and proper acknowledgement. While the preceding analysis has been focused on painting and visual art, future research is planned that will involve the analysis of additional fields of art including sound and music as well as art forms related to touch and smell. Augmented reality and virtual reality can also be explored and research into all these areas as they are related to artistic creativity and need to address the issue of the convergence of technologies and what issues arise when all of these subjects converge in the creation of artworks.

References

Alabama Policy Institute. (2020, November). Understanding the Difference Between Positive and Negative Rights. Alabama Policy Institute. https://alabamapolicy.org/wp-content/uploads/2020/11/GTI-Brief-Positive-Negative-Rights-1-1.pdf

Association for Computing Machinery (2018). ACM Code of Ethics and Professional Conduct. [online] Association for Computing Machinery. Available at: https://www.acm.org/code-of-ethics

Brey, Philip, A. E,, Anticipating Ethical Issues in emerging IT, Ethics and Information Technology (2012) 14:305-307

Brodie, N. (2019). The 'Art World' of the Auction Houses: The Role of Professional Experts. Arts, 8(2), 56. https://doi.org/10.3390/arts8020056

Butler, A. (2024). How to use Prisma App to Turn Your Photos into Works of Art. *Mobiography.* https://www.mobiography.net/apps/prisma-app/#:~:text=Prisma%20is%20a%20photo%20editing,of%20conventional%20photo%20editing%20apps

Caldwell, M. (2023). What is an "Author"? - Copyright Authorship of AI Art Through a Philosophical Lens. *Houston Law Review.* https://houstonlawreview.org/article/92132-what-is-an-author-copyright-authorship-of-ai-art-through-a-philosophical-lens

Cohn, Gabe. (2018). AI Art at Christie's Sells for $432,500. The New York Times. https://nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html

Ciulla, J.B., Martin, C.W. and Solomon, R.C. (2013). 3rd edn. Honest Work: A Business Ethics Reader. New York; Oxford: Oxford University Press.

Dafoe, T. (2023, February 16). A Class Action Lawsuit Against a Popular A.I. Art Generator Alleges the App Collects Its Users' Biometric Information Without Their Permission. Artnet News. https://news.artnet.com/art-world/class-action-lawsuit-lensa-ai-prisma-labs-biometric-information-2257096

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A Neural Algorithm of Artistic Style. *arXiv preprint*. doi: https://doi.org/10.48550/arXiv.1508.06576

Hong, E. (2024). Abstract cubist reinterpretations of Mona Lisa. [Image generated by DALL-E 3]. OpenAI.

Hong,E. (2024). Orca dancing in the ocean, reimagined in the style of Van Gogh. [Image generated by DALL-E 3]. Open AI.

Jee, C. (2018, October 26). A controversial artwork created by AI has hauled in $435,000 at auction. MIT Technology Review. https://www.technologyreview.com/2018/10/26/139292/a-controversial-artwork-created-by-ai-has-hauled-in-435000-at-auction/#:~:text=A%20controversial%20artwork%20created%20by%20AI%20has%20hauled%20in%20%24435%2C000%20at%20auction

Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., & Song, M. (2020). Neural Style Transfer: A Review. *IEEE Transactions on Visualization and Computer Graphics*, 26(11), 3365-3385. doi: 10.1109/TVCG.2019.2921336.

Johnson, Deborah, "Technology with No Human Responsibility" May 22 2014 , Journal of Business Ethics

Kiama Art Gallery. (2015, May 5). Impressionism – The influence of Photography. Kiama Art Gallery. https://kiamaartgallery.wordpress.com/2015/05/05/impressionism-the-influence-of-photography/

Kinsella, E. (2018). The First AI-Generated Portrait Ever Sold at Auction Shatters Expectations, Fetching $432,500—43 Times Its Estimate. Artnet. https://news.artnet.com/market/first-ever-artificial-intelligence-portrait-painting-sells-at-christies-1379902

Miller, A. I. (2019). The artist in the machine: The world of AI-powered creativity. *The MIT Press.* https://doi.org/10.7551/mitpress/11585.001.0001

Miller, K.W. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. IT Professional, 13(3), pp.57–59. doi:https://doi.org/10.1109/mitp.2011.46.

Mohan,N. (2018). Combining content of a Mona Lisa image with style of a Cubist like image using VGG16 network. Neural Style Transfer. Linkedin. https://www.linkedin.com/pulse/artificial-intelligence-applications-neural-style-transfer-mohan

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). *An Unsupervised Machine Learning Algorithms: Comprehensive Review.* International Journal of Computing and Digital Systems, 13(1). http://dx.doi.org/10.12785/ijcds/130172

Nestor, M.W. and Wilson, R.L. (2022). Anticipatory ethics and the use of CRISPR in humans. Cham: Springer.

Norman, D. A. (2004). Emotional design: Why we love (or hate) everyday things. *Basic Books/Hachette Book Group.*

OpenAI. (2024). DALL-E 3. OpenAI. https://openai.com/index/dall-e-3/

Picasso, P. (1937). Weeping Woman [Oil on canvas]. Tate, London. https://www.tate.org.uk/art/artworks/picasso-weeping-woman-t05010

Somers, M. (2020, July 21). Deepfakes, explained. MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

Stanford Encyclopedia of Philosophy. (1996, September 22). Distributive Justice. Stanford University. https://plato.stanford.edu/entries/justice-distributive/

Tavani, H.T. (2009). Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology. Hoboken, N.J.: Wiley

Tiku, N. (2022, September 28). AI can now create any image in seconds, bringing wonder and danger. The Washington Post. https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/

Play my Thesis: Opportunities and Limitations of Building Serious Games With AI

Niklas Humble
Department of Information Technology, Uppsala University, Sweden

niklas.humble@it.uu.se

Abstract: The field of Artificial Intelligence (AI) has gained increased attention since the release of ChatGPT in late 2022. Following the popularity and wide application, several generative AI (GAI or GenAI) tools have been released, with capabilities of generating novel content such as text, images, audio, and video. Previous research has noted both opportunities and limitations with GAI for various fields. One field with high potential impact is the game industry and the subfield of serious games, where the purpose of the game extends that of pure entertainment. GAI could play an important role for design and development of serious games, where design teams are typically smaller (compared with big commercial games), and competences in game mechanics, graphics and other game resources can be limited. This paper explores the opportunities and limitations of GAI tools to support development of serious games. This is done through the development of a serious game, based on the author's PhD thesis, where various GAI tools are used to generate game content such as in-game dialogue, graphics, and audio. Development of the game artefact, which is named 'Computer Programming in Schools', follows the process of a design science research (DSR) project where emphasis is placed on the design in one iteration of development. The author's experiences of utilizing GAI tools during the development of the game were recorded in a researcher diary, together with screenshots and images from the game. In this paper, these experiences are discussed and compared to related research in order to seek the most salient opportunities and limitations for GAI to support design and development of serious games. The paper provides several hands-on examples of design and development of a serious game with GAI tools and concludes with a set of recommendations for the future use of GAI in game development.

Keywords: Artificial intelligence, Generative AI, Serious games, Game development, Design science

1. Introduction

The field of Artificial Intelligence (AI) has a long history, with the concept coined in the 1950s, but the idea can of course be traced back even further. With the release of ChatGPT in November of 2022 the field has gained increased attention by both scholars and industry professionals. Following the wide application of ChatGPT, several other AI tools have been developed and released to the public under the umbrella term Generative AI (GenAI or GAI for short). GAI include AI models and tools that can produce high quality and novel text, images, video, audio, and animation.

Previous research has raised questions of challenges and opportunities with GAI, for example potential impact in medicine, on schools, and higher education. An interesting field for the application of GAI is game development and especially the development of serious games. Serious games refer to games which are developed with a purpose beyond entertainment only, such as education. Contrary to big commercial games, serious games are often developed by smaller teams or the practitioners themselves. Challenges associated with this can be the difficulty of building a game (mechanics, resources, design and so forth), but also connecting game elements to learning outcomes.

The aim of this paper is to explore potential opportunities and limitations of GAI tools to support development of serious games. The structure follows the process of design science research (DSR) where a game, based on the author's PhD thesis (Humble 2023), is designed and developed with GAI tools to highlight the opportunities and limitations of utilizing these tools in practice. Figure 1 depicts the start menu of the game developed in this project, 'Computer Programming in Schools'. In this paper, the following questions are explored and answered: 1) What are potential opportunities with artificial intelligence for development of serious games? 2) What are potential limitations with artificial intelligence for development of serious games?



Figure 1: Start menu of 'Computer Programming in Schools'

2. Background

The concept of Artificial Intelligence (AI) was coined by John McCarthy in 1956 but the idea of intelligent machines had been around before, for example in Sci Fi writer Isaac Asimov's "Runaround" and the seminal paper "Computing Machinery and Intelligence" by Alan Turing. A common distinction in AI is that between Artificial General Intelligence (AGI) and narrow AI, where AGI refers to universal problem-solving capabilities and narrow AI to problem-solving in narrowed fields.

Since the release of ChatGPT in November of 2022 and the attention it and similar AI systems have raised in academia and industry, another concept has gained popularity: Generative Artificial Intelligence (GAI). GAI is an umbrella term for AI systems with the capability of creating, or generating, a variety of new content in the form of audio, programming code, images, text and so forth.

Research have highlighted that GAI technologies have the potential of transforming both science and research, for example, to write essays, summarize literature, identify research gaps, and conduct peer reviews (Van Dis et al. 2023). On the other hand, it has been noted that we should be careful of what work is passed over to AI since. As expressed in an article, following the release of ChatGPT:

> *"scientific record is ultimately one of the human endeavor of struggling with important questions. Machines play an important role, but as tools ... the product must come from—and be expressed by—the wonderful computer in our heads" (Thorp 2023).*

2.1 Generative AI and the Game Industry

Research has noted the potential impact of generative AI on the game industry. One study highlights that early career game developers had mixed feelings about integrating GAI tools in their work, including for example concerns about ethics, professional change, efficiency, and usability (Boucher et al. 2024). There are of course many potential use cases for GAI in game industry, such as 3D modelling and more natural interaction with non-playable characters (NPC), which could aid and enhance design and development. However, research has also noted the potential impact on occupations related to game industry, such as content creation, development, legal, and media due to generative AI, where some jobs could be partially or fully displaced (Zarifhonarvar 2023).

Prior research has also noted the potentials of generative AI for development of serious games, where the purposes go beyond that of pure entertainment. Design and development of serious games can be both complex and expensive due to the different expertise needed in game development (Colado et al. 2023). However, it has been highlighted that generative AI can, for example, support in serious game prototyping and development by automating and reducing some of the tasks that needs to be conducted, such as artistic content, graphical resources, and creation of scenarios and characters (Colado et al. 2023).

2.2 AI Generated Text

The most famous example of a GAI system focusing on text creation is ChatGPT released by OpenAI in November of 2022. These types of systems are often labelled as Large Language Models (LLMs) or ChatBots and both OpenAI's ChatGPT, Microsoft's Copilot, and Google's Gemini could be mentioned as examples. It has been highlighted that these types of tools could support in speeding up writing tasks, such scientific paper and grant applications, but that oversight by a human is advised (Jo 2023). Prior research has highlighted that GAI tools for text creation have difficulties in both understanding the details of the instructions and getting the details correct in the answers. For example, not being able to provide the correct number of words in a paragraph (Figure 2) or 'lying' about the content (Humble et al. 2023). Seemingly lying about content or outputting information that does not correspond with reality is a well-established problem with GAI and is often labelled as 'Artificial Hallucination'.

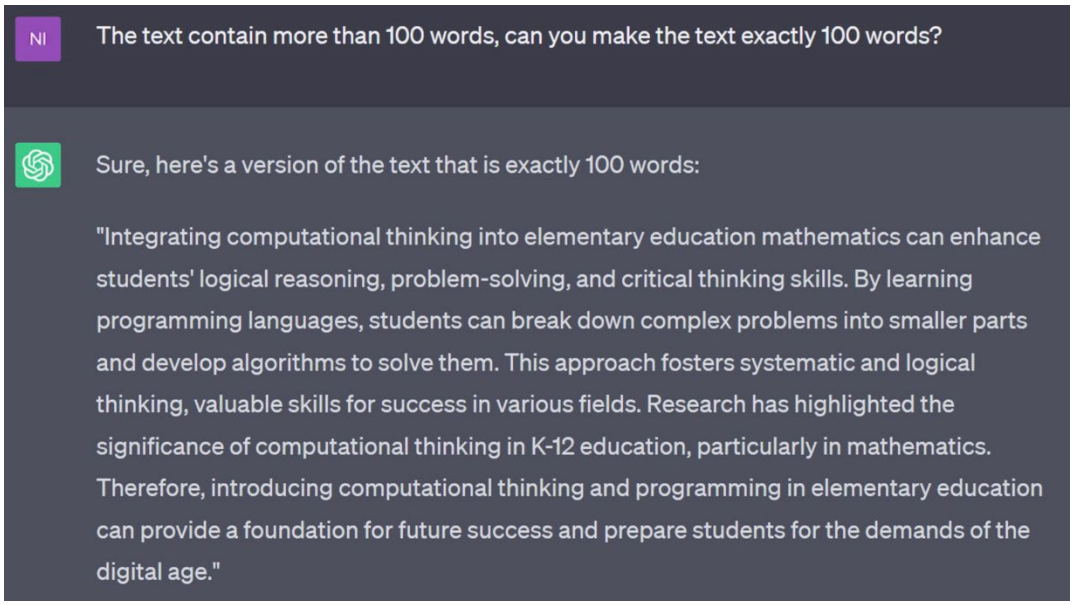


Figure 2: ChatGPT claiming that a paragraph is 100 words when it is not

2.3 AI Generated Images

Today, there are GAI systems that can produce high quality and realistic images in a range of different styles, using text prompts that describes the desired image to be generated. Some examples of text-to-image generative AI are DALL-E, MidJourney, StableDiffusion, and more recently Sora, an upcoming GAI system with the capability of generating realistic videos based on text prompts from the user. It has been highlighted that generative text-to-image AI brings with it both challenges and opportunities for the artistic field. It can foster innovation and creativity by making it more accessible, but also potentially threaten the creative job market (Vartiainen & Tedre 2023). GAI has also been discussed in relation to the usage of training data, and the potential ethical and legal aspects of using artists' and copyright holders' material without consent for developing these GAI models. Although the capabilities of many GAI tools to create realistic images are often impressive, there are limitations. Many have seen how GAI struggles with getting the correct number of fingers, hands, and legs when making images of humans. Similarly, GAI can struggle with a creative task that most elementary school children can master relatively quick, such as drawing a duck with basic shapes using the Python programming library Turtle (Figure 3) (Humble et al. 2023).

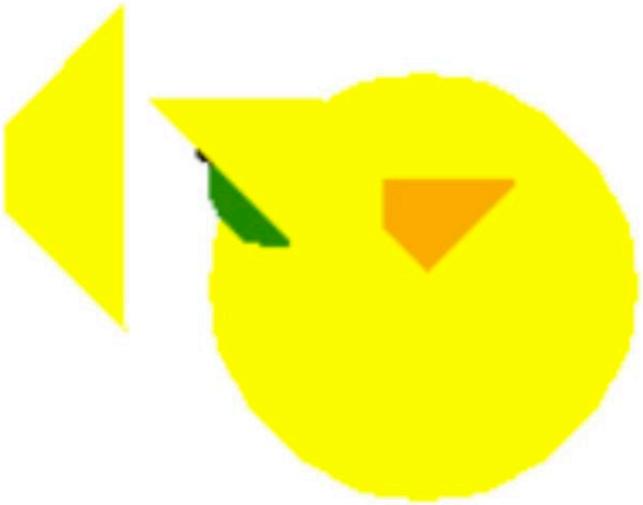


Figure 3: ChatGPT using Python library Turtle to draw a duck

2.4 AI Generated Audio

Another aspect of AI generated content is sound, through text-to-audio or text-to-music prompting. These GAI systems have been noted for their ability to, for example, assisting visually impaired people in their daily lives, providing text-to-speech, or image-to-speech, descriptions of everyday objects and facial expressions. However, it has been noted that there is a lack of diversity in generated audio and that it tends to sound homogeneous, especially within a specific audio category (Xie et al. 2024). Developing models for text-to-music generation also presents a specific challenge, compared to other type of audio generating, because of limited access to music that is not copyright protected and can be used for training (Chen et al. 2024).

3. Building a Serious Game with AI

Inspired by the activities in design science research (DSR) - which commonly include analysis and definition of a problem and requirements, as well as development and evaluation of an artefact - this paper seeks to explore the possibilities and limitations of recent AI technologies for game development, through the development of a serious game artefact based on the author's PhD thesis (Humble 2023). Since this paper will report on one iteration of development and focus mainly on design and development, the approach could be described as non-iterative (Figure 4) and design-focused (Figure 5) design science research (Humble & Mozelius 2023).

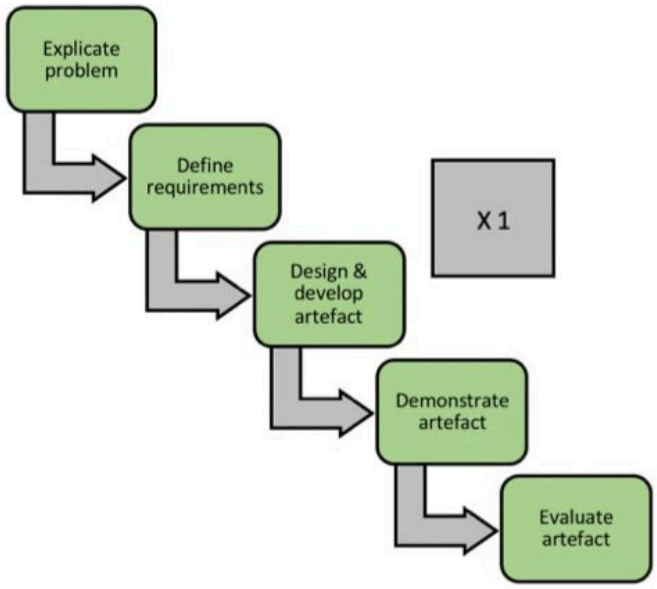


Figure 4: Non-iterative DSR

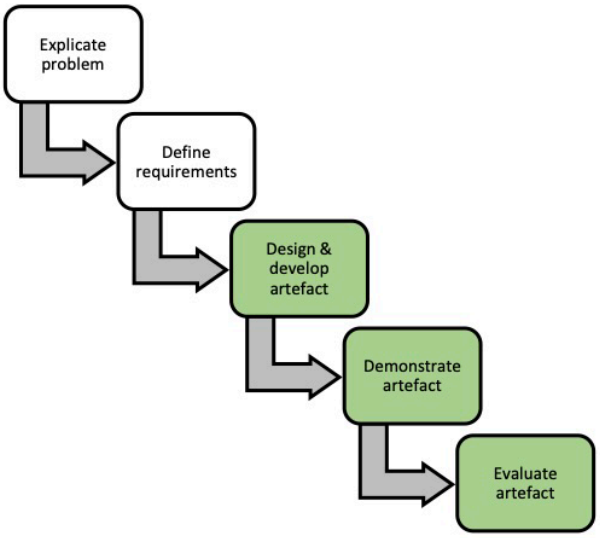


Figure 5: Design-focused DSR

Development of the game was conducted mainly during the fall semester of 2023, following the defense and publication of the PhD thesis in spring semester of 2023. During the design and development of the game artefact, which is labelled 'Computer Programming in Schools', a diary was kept to document tools that were used, thoughts about these tools, and reflections on the progress of the game development. The game has also been tested and informally evaluated by research colleagues and other contacts to the author, but no systematic data collection from potential end users have been conducted for evaluation of the game artefact. Testing and evaluation with potential end users should be part of future research and a second iteration of re-design and development.

The game artefact was developed using Ren'Py (www.renpy.org), a game engine specialized for visual novel games, which uses the Python programming language at its core. Game development is conducted by writing game scripts in a code editor (Figure 6) and adding graphics and audio for characters and scenery (Figure 7). Several AI tools were used to support design of in-game graphics, audio, and text. A combination of OpenAI's ChatGPT and Google's Gemini (previously Bard) were used to transform the content of the PhD thesis into in-game dialogue. Game graphics were generated using Microsoft Image Creator (DALL-E), and audio through BeatBot (www.beatbot.fm).

```
# The script of the game goes in this file.

# Declare characters used by this game. The color argument colorizes the
# name of the character.

define e = Character("Eileen")


# The game starts here.

label start:

    # Show a background. This uses a placeholder by default, but you can
    # add a file (named either "bg room.png" or "bg room.jpg") to the
    # images directory to show it.

    scene bg room

    # This shows a character sprite. A placeholder is used, but you can
    # replace it by adding a file named "eileen happy.png" to the images
    # directory.

    show eileen happy

    # These display lines of dialogue.

    e "You've created a new Ren'Py game."

    e "Once you add a story, pictures, and music, you can release it to the world!"

    # This ends the game.

    return
```

Figure 6: Default Ren'Py script in Visual Studio Code

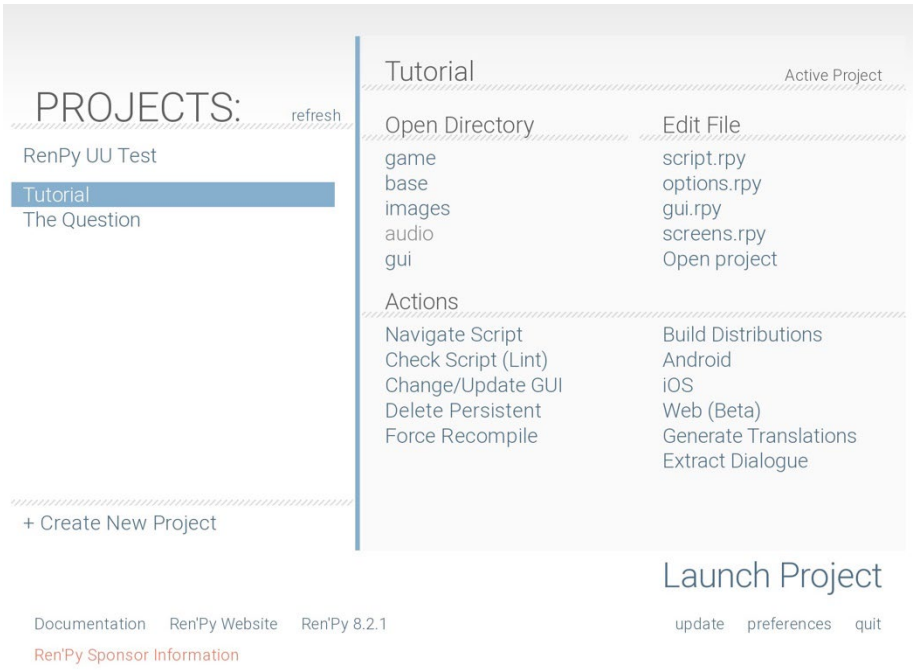


Figure 7: Main menu in the Ren'Py developer environment

4. AI Opportunities and Limitations for Game Development

In this section, design and development of the game artefact is presented and discussed. The discussion highlights opportunities and limitations of various AI tools that have been identified through development. These are presented related to the design of game dialogue, game graphics and game music in subheadings below.

4.1 Game Dialogue

The dialogue in the game was generated through the use of ChatGPT and Gemini. Text from the PhD thesis was provided to the AI tools together with instructions to summarize the text and make it more accessible to an audience outside of academia. Compared to what has been highlighted in previous research, that GAI tools can have the opportunity of speeding up the writing process in for example grant application and scientific papers (Jo 2023), holds true to some extent. Typically, ChatGPT or Gemini (or any other GAI tool) will not provide a 'perfect' answer on the first attempt. It was discovered early on in the work on the dialogue of the game artefact that both ChatGPT and Gemini works best when used in multiple iterations where the prompts are specified and refined. In that sense, quite a lot of work is needed to get the desired generated content. Therefore, not only human oversight is advised (Jo 2023), but also a clear goal of what the tool is to provide.

Providing the AI tools with more context usually generated better answers. For example, in the game artefact there is a cartoon cat as character (described further in the game graphics subheading) which can be included as information, together with the targeted audience, to set the tone of generated text. Through iterations and refinements of the prompts it was discovered that getting the generated text 'just as intended' was quite difficult and that it usually was better to aim for generating 'good enough' material that can be manually adjusted. For example, the targeted audience for the game artefact is not necessarily primary school children but adding that as information was usually a good strategy for making the language and choice of words less formal and less academic. The generated text could then be adjusted to suit the game dialogue and a wider audience, for example by removing content such as:

> *Prompt: "Summarize this text as if you were an enthusiastic cartoon cat with primary school children as the audience [...]"*
>
> *ChatGPT: "Hey there, kiddos! Imagine I'm your super-friendly cartoon cat, and I'm here to make this text super-duper fun and easy to understand! [...]"*

The biggest challenge for the AI tools regarding transforming the content of the PhD thesis to in-game dialogue was long, complex and multilayered texts. This can also be difficult for humans to understand so it should not

be considered surprising. Examples that often proved to be too difficult for the AI tools to handle sufficiently was the methodology section and the discussion section of the thesis (Figure 8). The author does not believe that this was because the content was difficult, but rather because there are many layers of information to keep track of in these sections. A common mistake for the AI generated content was to get the different layers of information wrong, for example mixing up findings of the thesis with that of related research. This could be compared to the concept of 'Artificial Hallucination'. Although, calling it 'lying' or 'hallucination' is a bit misguided in this case, but it is still a limitation of accuracy in the generated output that one should be mindful of when using GAI tools.

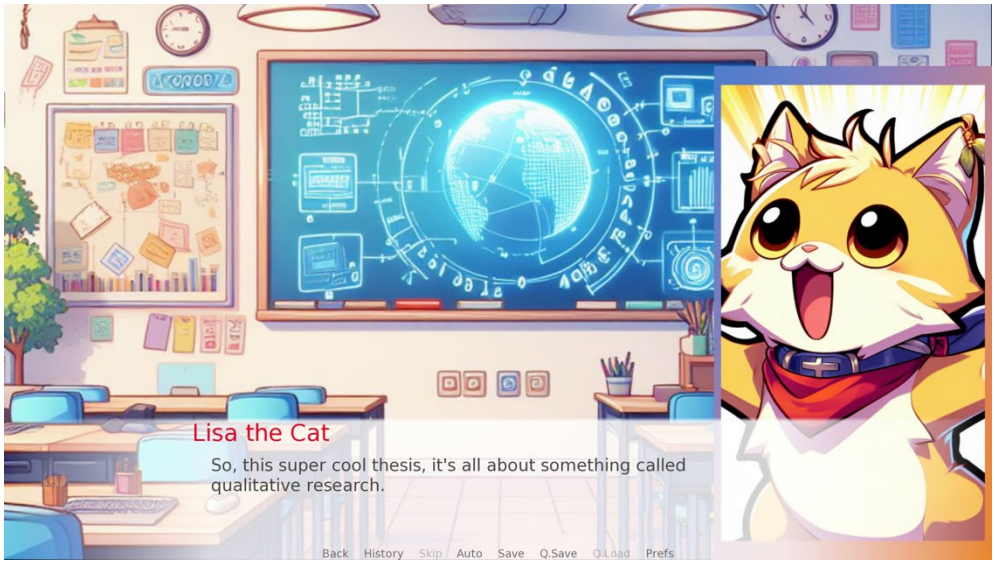


Figure 8: Screenshot from the game covering the methodology section of the thesis

4.2 Game Graphics

The graphics in the game were generated with Microsoft's Image Creator (DALL-E). The graphical style of the game is inspired by the game engine used for development, Ren'Py, which is typically used for developing visual novel games, and the context for the PhD thesis that the game content is based on, teachers' use of programming in secondary school mathematics and technology subjects. As previously highlighted, generative AI tools can produce high quality and realistic images in a range of different styles. However, during development it became evident that there are some limitations when using generative AI for image creation. For example, generative AI seems to be better at certain artistic styles, which is probably due to the access of training data.

Generative AI tools also seem to have serious limitations when it comes to the number of fingers, hands, arms, legs and so forth that should be depicted in an image. This affected the visual representations of the characters in the game, and to limit the potential mistakes in the graphical design there is only one human character, which is based on the author (Figure 9). The other two characters are depicted as animals inspired by programming language Python and the Scratch block programming environment, their names (Peter the Python and Lisa the Cat) were also chosen as an homage to two of the author's colleagues and friends from the time as a PhD student. Still, there were challenges when generating the images for the design. Peter the Python got more than one snake tail in several attempts of generating images, and on multiple occasion text appeared in the images with words and letters vaguely resembling real ones.

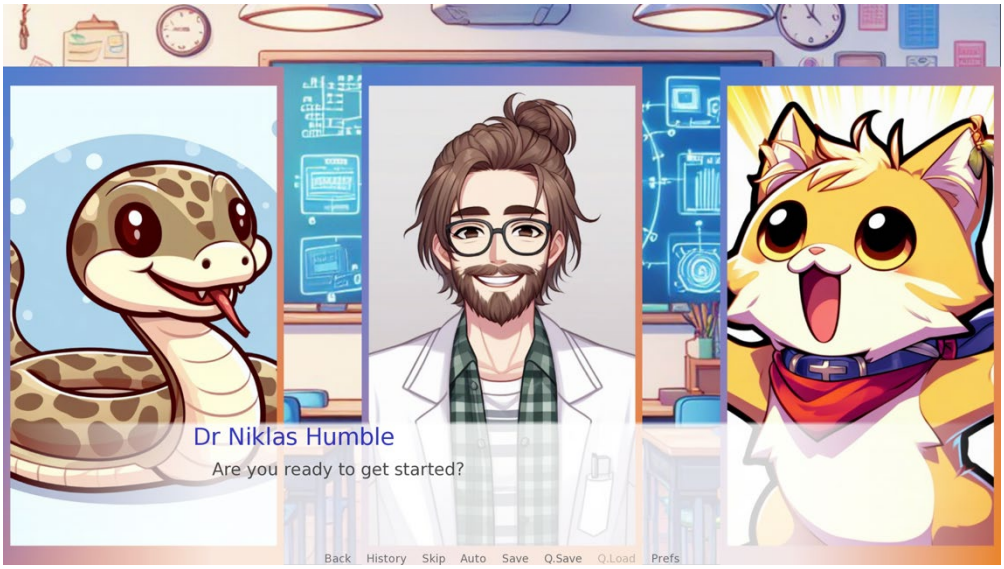


Figure 9: The characters in 'Computer Programming in Schools'

Previous research has highlighted both opportunities and challenges for the artistic field with generative AI, such as accessibility and creativity but also potential threat to the job market (Vartiainen & Tedre, 2023). In development of 'Computer Programming in Schools', both of these challenges and opportunities are present. Generative AI facilitates creativity by making game graphics more accessible to those less experienced. This is evident since the author would have had to collaborate with a graphical designer to make the graphics and artwork for the game if not for generative AI. Therefore, this is also a loosed work opportunity for graphical game designers. The potential ethical and legal aspects of using generative AI should also be considered, since there have been concerns raised that training data could include material from artists and copyright holders that have not given their consent.

4.3 Game Music

Music in the game were generated through BeatBot (www.beatbot.fm). In this tool, the user can do text-to-music prompting to generate the desired style of music. The style of the music was inspired by the graphical design in the game, which resembles Japanese Cartoon or Anime, and includes AI generated vocals singing about computer programming. As previously noted, there is a possibility of using AI systems to provide text-to-speech or image-to-speech assistance. Although this game, 'Computer Programming in Schools', did not use AI generated audio for text-to-speech or image-to-speech description, it is an obvious advantage for future game design, which could be incorporated in the second design cycle of the game to make it more accessible.

However, it should be noted that generating the intended style of music for the game was not a straightforward task. It was difficult to know exactly which words in the prompt that would have the desired outcome, therefore there was a lot of trial and error in testing different phrases. There was also a fair amount of luck involved in getting the desired result, since the same prompt could generate very different styles of music. Generating music and audio is limited by the AI models and their training data. Prior research has highlighted that generated audio tends to lack in diversity and often sound homogeneous (Xie et al. 2024). This was evident in development of 'Computer Programming in Schools', where certain genres of music that were tested had very limited diversity and different phrases in the prompts seemed to have little impact on the outcome (within that specific genre). It has also been highlighted regarding training data for AI generated music, as with other AI generated content, that there are potential ethical and legal issues to be considered (Chen et al. 2024).

5. Concluding Discussion

As this paper has shown, there are several opportunities with AI for game development and serious games in particular (summary in Table 1). AI and GAI tools can support game development by speeding up tasks such writing story and dialogues and creating graphics and in-game audio. It can also foster creativity among inexperienced game developers and in small-scale projects by lowering the threshold for many aspects of game design, such as creative writing, sound design and artwork. Implementing more advanced AI technologies in the design process, such as text-to-speech and image-to-speech, can also aid in making games more accessible, for example for visually impaired people.

However, this paper has also highlighted several limitations or challenges with AI for game development (summary in Table 1). GAI technologies tend to have limitations in accuracy and details of generated content. For example, text can lie about the content (Artificial Hallucination) and images can produce the wrong numbers of arms, fingers, legs and so on. Some generated content can also have a lack of diversity, which is related to the (limited) access of training data for the AI models. Writing prompts for GAI tools is often not a straightforward task but requires a lot of trial and error and luck, which could be viewed as a limitation of the technology. Although many of the present AI technologies comes short compared to professionals working with game design, the risks of reduced work opportunities are still obvious because of accessibility and possibility of time efficiency with these tools. It should also be noted that there could potentially be ethical and legal issues related to the training data, for example copyright, which the AI models are based on.

Table 1: Summary of opportunities and limitations with AI for game development

Opportunities	**Limitations**
Speed and efficiency	Accuracy and details
Creativity	Diversity
Support and accessibility	Jobs
	Trial and error and luck
	Ethics and law

Based on the opportunities and limitations with AI for game development that have been addressed in this paper, the author would like to give the following recommendations:

1. **Use AI for support and not to replace** – AI can speed up many tasks but often lacks in details and accuracy.
2. **Use AI for simple and narrow tasks** – AI can provide efficiency and foster creativity but is still limited by its training data.
3. **Use AI mainly for prototyping** – AI can generate realistic and impressive content but the quality can be varied and there could be potential ethical and legal issues with training data.

Lastly, it should be mentioned that efforts to identify opportunities and limitations with a fast-developing field, such as AI, is like trying to hit a moving target. Some of the opportunities and limitations that have been addressed in this paper will probably be perceived as outdated by the time of publication. Therefore, it is important that continuous research efforts are being made to study the implications of AI on game design and serious games. Especially important should be the investigation and development of ethical and legal practices concerning the use and development of generative AI.

References

Boucher, J. D., Smith, G., & Telliel, Y. D. (2024, May). Is Resistance Futile?: Early Career Game Developers, Generative AI, and Ethical Skepticism. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-13). https://doi.org/10.1145/3613904.3641889

Chen, K., Wu, Y., Liu, H., Nezhurina, M., Berg-Kirkpatrick, T., & Dubnov, S. (2024, April). MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1206-1210). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10447265

Colado, I. J. P., Colado, V. M. P., Morata, A. C., Píriz, R. S. C., & Manjón, B. F. (2023, October). Using new AI-driven techniques to ease serious games authoring. In *2023 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE. https://doi.org/10.1109/FIE58773.2023.10343021

Humble, N. (2023). *Programming in grade 7-9: Action possibilities and constraints from the perspective of mathematics and technology teachers*. Doctoral dissertation, Mid Sweden University. https://urn.kb.se/resolve?urn=urn:nbn:se:hig:diva-40787

Humble, N., Boustedt, J., Holmgren, H., Milutinovic, G., Seipel, S., & Östberg, A. S. (2023). Cheaters or AI-Enhanced Learners: Consequences of ChatGPT for Programming Education. *Electronic Journal of e-Learning*. https://doi.org/10.34190/ejel.21.5.3154

Humble, N., & Mozelius, P. (2023). Design science for Small Scale Studies: Recommendations for Undergraduates and Junior Researchers. In *European Conference on Research Methodology for Business and Management Studies (ECRM 2023)* (pp. 87-92). https://doi.org/10.34190/ecrm.22.1.1702

Jo, A. (2023). The promise and peril of generative AI. *Nature, 614*(1), 214-216. https://doi.org/10.1038/d41586-023-00340-6

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science, 379*(6630), 313-313. https://doi.org/10.1126/science.adg7879

Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature, 614*(7947), 224-226. https://doi.org/10.1038/d41586-023-00288-7

Vartiainen, H., & Tedre, M. (2023). Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity, 34*(1), 1-21. https://doi.org/10.1080/14626268.2023.2174557

Xie, Z., Li, B., Xu, X., Wu, M., & Yu, K. (2024, April). Enhancing Audio Generation Diversity with Visual Information. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 866-870). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10447384

Zarifhonarvar, A. (2023). Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Journal of Electronic Business & Digital Economics*. https://doi.org/10.1108/JEBDE-10-2023-0021

Generative Artificial Intelligence and the Impact on Sustainability

Niklas Humble[1] and Peter Mozelius[2,3]
[1]Uppsala University, Uppsala, Sweden
[2]University of Gävle, Gävle, Sweden
[3]Mid Sweden University, Östersund, Sweden

niklas.humble@it.uu.se
peter.mozelius@miun.se

Abstract: An increasingly popular subcategory of Artificial Intelligence (AI) is Generative AI (GAI), which encompasses technologies capable of creating new content, such as images, text, and music, often resembling outputs made by humans. The potential impact by GAI on sustainability is multifaceted. On the positive side, generative AI can aid in optimizing processes, developing innovative solutions, and identifying patterns in large datasets related to sustainability. This can lead to more efficient resource management, reduced energy consumption, and the creation of more sustainable products. However, there are also potential negative impacts, such as increased energy consumption associated with training and running generative AI models, as well as the potential for unintended consequences or biases in the generated content. Additionally, overreliance on generative AI may lead to reduced human oversight, which could undermine holistic, interdisciplinary, and collaborative approaches to sustainability. The aim of this paper is to explore the potential impacts on sustainability by generative artificial intelligence through a review of prior research on the topic. The study was conducted with a scoping literature review approach to identify potential impacts by generative AI on sustainability. Data were collected through a search in the database Scopus during the spring semester of 2024. Keywords, relevant for the study, were combined with Boolean operators. Papers identified through the search underwent a manual screening process by the authors, in which papers were selected for inclusion or exclusion in the study based on a set of criteria. Included paper were then analyzed with thematic analysis, according to the guidelines by Braun and Clarke. A categorization matrix, based in prior research on sustainability, supported the analysis and deductive coding of collected data. Results of the study highlight generative AI's potential impact on sustainability that relate to both environmental aspects, economic aspects, and social aspects of sustainability. These different aspects of sustainability impact make this research an important contribution for deepening the understanding of generative AI and its potential consequences for society. Findings of the study provide theoretical contribution, implications for practice, and recommendations for future research on generative AI and sustainability.

Keywords: Generative AI, Sustainability impact, Environmental sustainability, Economic sustainability, Social sustainability

1. Introduction

The release of ChatGPT in November 2022 by OpenAI has stimulated implementations of generative artificial intelligence (GAI) in various areas such as academia, business, healthcare and in society at large. AI has been a useful part of processes in numerous areas for several years, but the emergence of GAI applications such as ChatGPT, Gemini, and DALL-E could be considered a new AI spring with applications that are user-friendly and have a promising performance potential. (Kanbach et al, 2024) The umbrella term GAI encompasses technologies capable of creating content in different modalities, such as images, text, music and videos (Rabowsky, 2023), that often resembles content made by humans. The potential impact by GAI on individuals, organizations and sustainability could be described as a double-edged sword (Gursoy & Cai, 2024).

On the positive side, GAI can contribute to optimizing processes and identifying patterns in large datasets related to sustainability. This could result in more efficient resource management, reduced energy consumption, and the creation of more sustainable products. There are optimistic research studies claiming that GAI might contribute to biodiversity, water resource conservation, renewable energy, energy conservation and sustainable transportation systems (Nishant, Kennedy & Corbett, 2020), which could help companies lower their carbon footprint and preserve sensitive natural heritage sites and national parks. Moreover, AI may also be useful in combating the climate change by providing a more comprehensive understanding of climate change scenarios and carbon emission analyses (Gursoy & Cai, 2024).

On the negative side, there are also potential negative impacts, such as increased energy consumption associated with training and running generative AI models (de Vries, 2023), as well as the potential for unintended consequences or biases in the generated content (Shuford, 2024). Interacting with a GAI tool has been estimated to "likely cost 10 times more than a standard keyword search", and as a standard Google search has been reported to use 0.3 Wh of electricity, the electricity consumption of a GAI tool could be as high as approximately 3 Wh per interaction (de Vries, 2023). An example of unpleasant bias in GAI systems is the facial recognition technology used by law enforcement agencies. A research study found that facial recognition

technology showed significantly lower accuracy rates for individuals with darker skin tones. A bias that can lead to severe consequences such as wrongful arrests and convictions. (Shuford, 2024)

Finally, overreliance on AI may lead to reduced human oversight, which could undermine holistic, interdisciplinary, and collaborative approaches to sustainability. With this background, it is important that AI researchers and practitioners are aware of the potential impact that generative AI may have on sustainability. The aim of this paper is to explore the potential impacts on sustainability by generative AI through a literature review of prior research on the topic.

2. Theoretical Framework

Sustainability has been defined as the development that meets the needs of the present without compromising the ability of future generations (WCED, 1987; Gursoy & Cai, 2024). The concept of sustainability has been broadly acknowledged as multidimensional, the various dimensions have evolved over time and have often been treated as separate branches. One common representation of sustainability depicts it with the three sub concepts of environmental, economic, and social sustainability as intersecting circles (Purvis, Mao & Robinson, 2019). While yet another representation depicts the same three sub concepts as pillars holding up the overarching concept of Sustainability as the roof (Purvis, Mao & Robinson, 2019).

Although this paper uses the sustainability concepts and the sub concepts of environmental, economic, and social sustainability; it does not make any claim on which type of representation (pillars, circles or other) that is best suited for capturing the nuances of the concepts. The concepts of environmental, economic, and social sustainability are further described below.

The sustainability concept comprises the relationships between humans and nature with a concern on the global environmental problems. Two examples of global environmental problems are industrial pollution and climate change (Giovannoni & Fabietti, 2013). The environmental sustainability concept is not a completely isolated branch, but a concept that is related to economic sustainability. Ideas of a sustainable link between environment and business have been suggested several decades ago. With ecologically friendly strategies and green marketing as a way to attract the environmentally conscious consumers (Sharma et al., 2010).

The further alignment to environmental sustainability, recycling and circular economy are important components of Industry 4.0 (Tang et al., 2022). Economic sustainability also involves companies' and organizations' resilience to crisis, where the recent Covid-19 pandemic is an example of a global crisis with economic challenges. Moreover, the Covid-19 pandemic also has illustrated the relationship between economic sustainability and social sustainability, and how they are intertwined in a severe crisis. (Rai et al., 2021) In a circular economy, the economic growth should be aligned with social responsibility and a general enhancement of sustainability (Upadhyay et al., 2021).

The concept of social sustainability overlaps with corporate social responsibility in the concern of well-being for humans and society through a thoughtful management of social resources (Panigrahi et al., 2019). Other important main aspects of social sustainability are the alleviation of poverty, the respect for justice and human rights, employees' general welfare and work transparency (Kumar & Rahman, 2017; Govindan, Shaw & Majumdar, 2021). A suggested indicator of social sustainability is urban green space accessibility (Du & Zhang, 2020), which is one of several areas where social sustainability connects with environmental sustainability as well as with economic sustainability.

3. Method

The study was conducted as a scoping literature review to explore and map (Munn et al., 2018) the potential impacts on sustainability by GAI systems, as this has been highlighted in prior research. The processes of data collection and data analysis are described in subheadings below.

3.1 Data Collection

Data were collected through a search in the Scopus database in the Spring semester of 2024. The query string combined keywords Generative, Artificial Intelligence, and Sustainability in a search of papers' titles, abstracts, and keywords. The full query string used was:

> (TITLE-ABS-KEY (generative) AND TITLE-ABS-KEY (artificial AND intelligence) AND TITLE-ABS-KEY (sustainability))

The initial search identified 60 potentially relevant papers to be included in the study. Search filters were then applied which brought down the number of potentially relevant papers to 26. The search filters were: Open Access, English language, journal articles, and conference papers. The 26 papers then underwent full paper screening in which 4 papers were excluded for not addressing potential impact on sustainability by generative AI. This resulted in 22 papers being included in the study for further analysis (depicted in Table 2). A summary of inclusion and exclusion criteria are presented in Table 1.

Table 1: Inclusion and exclusion criteria for selected papers

Inclusion	Exclusion
Open Access	Not Open Access
English	Not English
Journal article or conference paper	Not journal article or conference paper
Address (potential) impact on sustainability by generative artificial intelligence	Does not address (potential) impact on sustainability by generative artificial intelligence

3.2 Data Analysis

Data analysis was conducted with a deductive thematic analysis and supported by the sustainability concept, with environmental, economic, and social sustainability as subcategories. After creating a spreadsheet categorization matrix, comprised by the sustainability subcategories, the analysis followed Braun & Clarke's (2006) six phases for thematic analysis.

1. Data familiarization, through paper screening and repeated reading.
2. Generate initial codes, by collecting relevant extracts from the data in the categorization matrix.
3. Searching for themes, by arranging and re-arranging codes, based on similarities and differences, in the categorization matrix.
4. Reviewing themes, by reconsidering arrangements, or clusters, of codes to create coherent patterns.
5. Defining and naming themes, based on the content in the codes and related research.
6. Producing the report, by structuring the presentation of themes according to the categorization matrix.

4. Results and Analysis

This section presents the results and analysis of the 22 included papers (Table 2) and discuss implications of the findings for sustainability. Findings are presented and discussed according to the three sub concepts of sustainability.

Table 2: Included papers

Number	Paper
1	Adiasto, K. (2024). SustAInable Employability: Sustainable Employability in the Age of Generative Artificial Intelligence. *Group & Organization Management*, 10596011241238792. https://doi.org/10.1177/10596011241238792
2	Almaz, A. F., El-Agouz, E. A. E., Abdelfatah, M. T., & Mohamed, I. R. (2024). The Future Role of Artificial Intelligence (AI) Design's Integration into Architectural and Interior Design Education is to Improve Efficiency, Sustainability, and Creativity. *Civil Engineering and Architecture, 12*(3), 1749–1772. https://doi.org/10.13189/cea.2024.120336
3	Alshahrani, A. (2023). The impact of ChatGPT on blended learning: Current trends and future research directions. *International Journal of Data and Network Science, 7*(4), 2029–2040. https://doi.org/10.5267/j.ijdns.2023.6.010
4	Aseeva, A. (2023). Liable and Sustainable by Design: A Toolbox for a Regulatory Compliant and Sustainable Tech. *Sustainability, 16*(1), 228. https://doi.org/10.3390/su16010228
5	Currie, G. M., Hawk, K. E., & Rohren, E. M. (2024). The potential role of artificial intelligence in sustainability of nuclear medicine. *Radiography*, S1078817424000671. https://doi.org/10.1016/j.radi.2024.03.005
6	De Villiers, C., Dimes, R., & Molinari, M. (2024). How will AI text generation and processing impact sustainability reporting? Critical analysis, a conceptual framework and avenues for future research. *Sustainability Accounting, Management and Policy Journal, 15*(1), 96–118. https://doi.org/10.1108/SAMPJ-02-2023-0097

Number	Paper
7	Eglash, R., Robert, L., Bennett, A., Robinson, K. P., Lachney, M., & Babbitt, W. (2020). Automation for the artisanal economy: Enhancing the economic and environmental sustainability of crafting professions with human–machine collaboration. *AI & SOCIETY, 35*(3), 595–609. https://doi.org/10.1007/s00146-019-00915-w
8	Emmert-Streib, F. (2023). What Is the Role of AI for Digital Twins? *AI, 4*(3), 721–728. https://doi.org/10.3390/ai4030038
9	Fathoni, A. F. C. A. (2023). Leveraging Generative AI Solutions in Art and Design Education: Bridging Sustainable Creativity and Fostering Academic Integrity for Innovative Society. *E3S Web of Conferences, 426*, 01102. https://doi.org/10.1051/e3sconf/202342601102
10	Gallent Torres, C., Zapata González, A., & Ortego Hernando, J. L. (2023). El impacto de la inteligencia artificial generativa en educación superior: Una mirada desde la ética y la integridad académica [The impact of Generative Artificial Intelligence in higher education: a focus on ethics and academic integrity]. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa, 29*(2). https://doi.org/10.30827/relieve.v29i2.29134
11	Haluza, D., & Jungwirth, D. (2023). Artificial Intelligence and Ten Societal Megatrends: An Exploratory Study Using GPT-3. *Systems, 11*(3), 120. https://doi.org/10.3390/systems11030120
12	Lin, Y. (2022). Rhetoric, Writing, and Anexact Architecture: The Experiment of Natural Language Processing (NLP) and Computer Vision (CV) in Architectural Design. *Proceedings of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*, 343–352. https://doi.org/10.52842/conf.caadria.2022.1.343
13	Rampini, L., & Cecconi, F. R. (2022). Artificial intelligence in construction asset management: A review of present status, challenges and future opportunities. *Journal of Information Technology in Construction, 27*, 884–913. https://doi.org/10.36680/j.itcon.2022.043
14	Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science, 19*(1), 27. https://doi.org/10.1186/s13012-024-01357-9
15	Richards, D., Worden, D., Song, X. P., & Lavorel, S. (2024). Harnessing generative artificial intelligence to support nature-based solutions. *People and Nature, 6*(2), 882–893. https://doi.org/10.1002/pan3.10622
16	Šedlbauer, J., Činčera, J., Slavík, M., & Hartlová, A. (2024). Students' reflections on their experience with CHATGPT. *Journal of Computer Assisted Learning, jcal*.12967. https://doi.org/10.1111/jcal.12967
17	Spiegelhalter, T., & Werner, L. C. (2022). Methods for In-Silico Environmental Resilience, 2018 to 2100. *IOP Conference Series: Earth and Environmental Science, 1078*(1), 012139. https://doi.org/10.1088/1755-1315/1078/1/012139
18	Tomar, P., & Grover, V. (2024). Transforming the Energy Sector: Addressing Key Challenges through Generative AI, Digital Twins, AI, Data Science and Analysis. *EAI Endorsed Transactions on Energy Web, 10*. https://doi.org/10.4108/ew.4825
19	Tortora, L. (2024). Beyond Discrimination: Generative AI Applications and Ethical Challenges in Forensic Psychiatry. *Frontiers in Psychiatry, 15*, 1346059. https://doi.org/10.3389/fpsyt.2024.1346059
20	Vidmar, M., Fleck, J., & Williams, R. (2023). AI and Data in Engineering and Innovation: Towards a Sustainable Future? *2023 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 1–2. https://doi.org/10.1109/ICE/ITMC58018.2023.10332336
21	Vimpari, V., Kultima, A., Hämäläinen, P., & Guckelsberger, C. (2023). "An Adapt-or-Die Type of Situation": Perception, Adoption, and Use of Text-to-Image-Generation AI by Game Industry Professionals. *Proceedings of the ACM on Human-Computer Interaction, 7(CHI PLAY)*, 131–164. https://doi.org/10.1145/3611025
22	Zang, Z., & Ding, W. (2024). Eco-Centric Generative Design Workflow: Extending Sustainability in Architecture. In F. Ying, L. C. Jain, R. Wan, Q. Wu, & F. Shi (Eds.), *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/FAIA240006

4.1 Environmental Impact

The analysis identified two themes related to environmental sustainability: Resources and Nature-based. The theme of Nature-based is supported by codes from Almaz et al. (2024), De Villiers et al. (2024), and Richards et al. (2024). The codes highlight potential positive impacts of GAI on environmental sustainability. For example, that GAI can be used for farm planning (Richards et al., 2024), identifying, and potentially finding solutions for, sustainability issues (De Villiers et al., 2024), and that GAI could be combined with biomimicry principles to create sustainable and visually stunning structures in architecture (Almaz et al., 2024).

The theme of Resources is supported by codes from Currie, Hawk and Rohren (2024), Tomar and Grover (2024), Tortora (2024), Zang and Ding (2024), Emmert-Streib (2023), Fathoni (2023), Haluza and Jungwirth (2023),

Vidmar, Fleck and Williams (2023), Lin (2022), Rampini and Cecconi (2022), and Spiegelhalter and Werner (2022). The codes highlight both potential positive impacts and potential negative impacts of GAI on environmental sustainability. For example, it is brought up that GAI could support in minimizing the use of resources (Fathoni, 2023; Haluza & Jungwirth, 2023; Rampini & Cecconi, 2022) and make energy sources, such as wind energy, more efficient (Emmert-Streib, 2023; Tomar & Grover, 2024). On the other hand, codes also note that GAI require energy to perform computational tasks (Currie, Hawk & Rohren, 2024) and that the training of these AI models require huge amounts of energy and physical infrastructure which is likely to continuing to increase (Tortora, 2024).

4.2 Economic Impact

The analysis identified one theme related to economic sustainability: Optimization. The theme is supported by codes from Almaz et al. (2024), Currie, Hawk and Rohren (2024), De Villiers et al. (2024), and Zang and Ding (2024). The codes highlight mainly potential positive impacts of GAI on economic sustainability. For example, that GAI can optimize architectural design by enabling the possibility of quickly generating multiple options for design, which can save both resources and time (Almaz et al., 2024; Zang & Ding, 2024). Further, it is noted that GAI could optimize accounting and management by providing organizations with improved information efficiency, accessibility, and accuracy, which would be a reduced burden to many companies (De Villiers et al., 2024). However, Currie, Hawk and Rohren (2024) emphasize that there is a delicate balance between potential economic (and sustainability) gains and human values, which becomes evident in contexts such as medicine.

4.3 Social Impact

The analysis identified four themes related to social sustainability: Understanding, Well-being, Trust and equity, and Sustainable workforce. The theme of Sustainable workforce is supported by codes from Adiasto (2024) and Currie, Hawk and Rohren (2024). The codes highlight mainly potential positive impacts of GAI on social sustainability. For example, that GAI could be used to prevent burnouts and relieve strained workforce of pressure by automating repetative and less cognitive demanding tasks, which could free up time for more human-centered tasks (Currie, Hawk & Rohren, 2024). However, it has also been noted that GAI could be a double-edged sword. Meaning that it could support in completing tasks faster, but that does not neccissarily imply that employes will get more free time. They could instead be asked to perform more tasks in the same time, and if so, will they be compensated with a higher salary? (Adiasto, 2024)

The theme of Trust and equity is supported by codes from Currie, Hawk and Rohren (2024), Aseeva (2023), and Vidmar, Fleck and Williams (2023). The codes highlight mainly potential negative impacts of GAI on social sustainability. For example, that GAI has issues with explainability due to its black box nature which undermines trust in the technology (Vidmar, Fleck & Williams, 2023). Further, it is noted that even though GAI could be used to address inequity and social asymmetry, it also has the potential of weding the gap of equity (Currie, Hawk & Rohren, 2024). The theme of Well-being is supported by codes from Currie, Hawk and Rohren (2024), Reddy (2024), Vimpari et al. (2023), and Lin (2022). The codes highlight potential positive impacts of GAI on social sustainability. For example, that GAI could, if responsibly introduced, enhance affordability, quality and accessibility in healthcare (Reddy, 2024), by supporting physicians with automated patient pre-readings and identifying potential urgent health issues which could save lives (Currie, Hawk and Rohren, 2024). GAI could also add to professionals' well-being by affording self-realization and artistic expression through co-creation between human and AI (Vimpari et al., 2023).

The theme of Understanding is supported by codes from De Villiers et al. (2024), Šedlbauer et al. (2024), Alshahrani (2023), Fathoni (2023), Gallent Torres et al. (2023), and Eglash et al. (2020). The codes highlight potential positive impacts of GAI on social sustainability. For example, that GAI could support understanding of sustainability (Šedlbauer et al., 2024), foster sustainable creativity (Fathoni, 2023), and engagement, motivation and self-directed learning through the use of AI chatbots (Alshahrani, 2023) in education. It is also noted that GAI could be used for empowerment and social justice (Eglash et al. (2020), and to increase engagement and understanding of company information and support organizations in communicating impact of sustainability initiatives (De Villiers et al., 2024).

5. Concluding Discussion

The findings presented in the previous section are summarized in Figure 1, which also shows how the themes align in a positive-negative impact continuum regarding GAI in relation to sustainability. The themes have been classified based on the frequency of positive or negative in the codes that aggregated the themes.

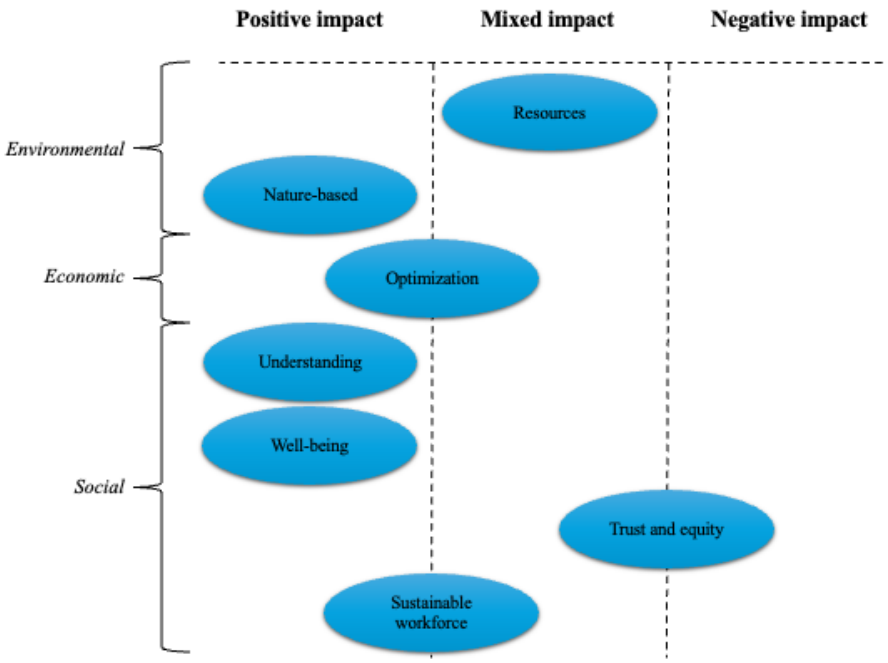


Figure 1: Found GAI themes alignment in a positive-negative impact continuum on sustainability

Despite that there were no exclusion criteria based on publication years, the entire result set were published within the past 4 years. Based on the analyzed result set it seems to have been a general positive attitude towards GAI's ability to handle sustainability issues during this period. This leads to the preliminary conclusion that the recent years has been a new spring for AI, and GAI in particular, which also has been highlighted in previous research (Olhede & Wolfe, 2018; Maclure, 2020). Moreover, there are also researchers that expects this third spring of AI to be followed by a third AI winter (Hajkowicz et al., 2023). To further confirm this, the study should be extended and more thoroughly conducted, for example through a systematic literature review.

Future research should also seek to balance the positive findings in this paper by exploring more of the negative impact of GAI to sustainability. Challenges such as energy consumption, physical infrastructure, and ethical and legal issues are important to consider in the exploration of GAI's capabilities. Further, authors of this paper would suggest that future research develop more specific and detailed definitions of GAI and associated technologies. This would allow for more accurate investigations of different AI technologies impact on sustainability.

It should be noted that this literature review was conducted with a limited approach, with limited number of papers and not including any tests of results, such as analysis of the evolution of publications, cluster analysis, analysis of future lines of research and so forth. The justification for this being to provide an initial exploration of potential GAI impacts on sustainability, which should be extended in future research. An important next step is therefore to include more databases in the literature search, such as Web of Science, to provide a more comprehensive dataset. Inclusion and exclusion criteria could also be revised to include, for example, papers that are not Open Access. This could potentially have negative impact on readers ability to reach included papers for themselves but would also contribute to a more comprehensive understanding of the research field. These adjustments to the ongoing research should also have a positive impact on the number of included papers.

To further widen and strengthen the literature review, alternative keywords for generative AI and sustainability should be added to the query string, as well as backward and forward searches based on the included papers. Tests of the results should be included as part of the methodology.

References

Adiasto, K. (2024). SustAInable Employability: Sustainable Employability in the Age of Generative Artificial Intelligence. *Group & Organization Management,* 10596011241238792. https://doi.org/10.1177/10596011241238792

Almaz, A. F., El-Agouz, E. A. E., Abdelfatah, M. T., & Mohamed, I. R. (2024). The Future Role of Artificial Intelligence (AI) Design's Integration into Architectural and Interior Design Education is to Improve Efficiency, Sustainability, and Creativity. *Civil Engineering and Architecture, 12*(3), 1749–1772. https://doi.org/10.13189/cea.2024.120336

Alshahrani, A. (2023). The impact of ChatGPT on blended learning: Current trends and future research directions. *International Journal of Data and Network Science, 7*(4), 2029–2040. https://doi.org/10.5267/j.ijdns.2023.6.010

Aseeva, A. (2023). Liable and Sustainable by Design: A Toolbox for a Regulatory Compliant and Sustainable Tech. *Sustainability, 16*(1), 228. https://doi.org/10.3390/su16010228

Ayoubi, K. (2024). Adopting ChatGPT: Pioneering a new era in learning platforms. *International Journal of Data and Network Science, 8*(2), 1341–1348. https://doi.org/10.5267/j.ijdns.2023.11.001

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Currie, G. M., Hawk, K. E., & Rohren, E. M. (2024). The potential role of artificial intelligence in sustainability of nuclear medicine. *Radiography*, S1078817424000671. https://doi.org/10.1016/j.radi.2024.03.005

De Villiers, C., Dimes, R., & Molinari, M. (2024). How will AI text generation and processing impact sustainability reporting? Critical analysis, a conceptual framework and avenues for future research. *Sustainability Accounting, Management and Policy Journal, 15*(1), 96–118. https://doi.org/10.1108/SAMPJ-02-2023-0097

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule, 7*(10), 2191-2194. https://doi.org/10.1016/j.joule.2023.09.004

Du, M., & Zhang, X. (2020). Urban greening: A new paradox of economic or social sustainability?. *Land Use Policy, 92*, 104487. https://doi.org/10.1016/j.landusepol.2020.104487

Eglash, R., Robert, L., Bennett, A., Robinson, K. P., Lachney, M., & Babbitt, W. (2020). Automation for the artisanal economy: Enhancing the economic and environmental sustainability of crafting professions with human–machine collaboration. *AI & SOCIETY, 35*(3), 595–609. https://doi.org/10.1007/s00146-019-00915-w

Emmert-Streib, F. (2023). What Is the Role of AI for Digital Twins? *AI, 4*(3), 721–728. https://doi.org/10.3390/ai4030038

Fathoni, A. F. C. A. (2023). Leveraging Generative AI Solutions in Art and Design Education: Bridging Sustainable Creativity and Fostering Academic Integrity for Innovative Society. *E3S Web of Conferences, 426*, 01102. https://doi.org/10.1051/e3sconf/202342601102

Gallent Torres, C., Zapata González, A., & Ortego Hernando, J. L. (2023). El impacto de la inteligencia artificial generativa en educación superior: Una mirada desde la ética y la integridad académica [The impact of Generative Artificial Intelligence in higher education: a focus on ethics and academic integrity]. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa, 29*(2). https://doi.org/10.30827/relieve.v29i2.29134

Giovannoni, E., & Fabietti, G. (2013). What is sustainability? A review of the concept and its applications. *Integrated reporting: Concepts and cases that redefine corporate accountability*, 21-40. https://doi.org/10.1007/978-3-319-02168-3_2

Govindan, K., Shaw, M., & Majumdar, A. (2021). Social sustainability tensions in multi-tier supply chain: A systematic literature review towards conceptual framework development. *Journal of cleaner production, 279*, 123075. https://doi.org/10.1016/j.jclepro.2020.123075

Gursoy, D., & Cai, R. (2024). Artificial intelligence: an overview of research trends and future directions. *International Journal of Contemporary Hospitality Management*. https://doi.org/10.1108/IJCHM-03-2024-0322

Hajkowicz, S., Sanderson, C., Karimi, S., Bratanova, A., & Naughtin, C. (2023). Artificial intelligence adoption in the physical sciences, natural sciences, life sciences, social sciences and the arts and humanities: A bibliometric analysis of research publications from 1960-2021. *Technology in Society*, *74*, 102260. https://doi.org/10.1016/j.techsoc.2023.102260

Haluza, D., & Jungwirth, D. (2023). Artificial Intelligence and Ten Societal Megatrends: An Exploratory Study Using GPT-3. *Systems, 11*(3), 120. https://doi.org/10.3390/systems11030120

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024). The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, *18*(4), 1189-1220. https://doi.org/10.1007/s11846-023-00696-z

Kumar, D., & Rahman, Z. (2017). Analyzing enablers of sustainable supply chain: ISM and fuzzy AHP approach. *Journal of Modelling in Management, 12*(3), 498-524. https://doi.org/10.1108/JM2-02-2016-0013

Lin, Y. (2022). Rhetoric, Writing, and Anexact Architecture: The Experiment of Natural Language Processing (NLP) and Computer Vision (CV) in Architectural Design. *Proceedings of the 27th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*, 343–352. https://doi.org/10.52842/conf.caadria.2022.1.343

Maclure, J. (2020). The new AI spring: a deflationary view. *AI & society*, *35*(3), 747-750. https://doi.org/10.1007/s00146-019-00912-z

Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC medical research methodology, 18*, 1-7. https://doi.org/10.1186/s12874-018-0611-x

Nishant, R., Kennedy, M., & Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *International Journal of Information Management*, *53*, 102104. https://doi.org/10.1016/j.ijinfomgt.2020.102104

Olhede, S., & Wolfe, P. (2018). The AI spring of 2018. *Significance*, *15*(3), 6-7. https://doi.org/10.1111/j.1740-9713.2018.01140.x

Panigrahi, S. S., Bahinipati, B., & Jain, V. (2019). Sustainable supply chain management: A review of literature and implications for future research. *Management of Environmental Quality: An International Journal*, *30*(5), 1001-1049. https://doi.org/10.1108/MEQ-01-2018-0003

Purvis, B., Mao, Y., & Robinson, D. (2019). Three pillars of sustainability: in search of conceptual origins. *Sustainability science, 14,* 681-695. https://doi.org/10.1007/s11625-018-0627-5

Rabowsky, B. (2023). Applications of Generative AI to Media. *SMPTE Motion Imaging Journal*, *132*(8), 53-57. http://dx.doi.org/10.5594/JMI.2023.3297238

Rai, S. S., Rai, S., & Singh, N. K. (2021). Organizational resilience and social-economic sustainability: COVID-19 perspective. *Environment, Development and Sustainability*, *23*, 12006-12023. https://doi.org/10.1007/s10668-020-01154-6

Rampini, L., & Cecconi, F. R. (2022). Artificial intelligence in construction asset management: A review of present status, challenges and future opportunities. *Journal of Information Technology in Construction, 27*, 884–913. https://doi.org/10.36680/j.itcon.2022.043

Reddy, S. (2024). Generative AI in healthcare: An implementation science informed translational path on application, integration and governance. *Implementation Science, 19*(1), 27. https://doi.org/10.1186/s13012-024-01357-9

Richards, D., Worden, D., Song, X. P., & Lavorel, S. (2024). Harnessing generative artificial intelligence to support nature-based solutions. *People and Nature, 6*(2), 882–893. https://doi.org/10.1002/pan3.10622

Šedlbauer, J., Činčera, J., Slavík, M., & Hartlová, A. (2024). Students' reflections on their experience with CHATGPT. *Journal of Computer Assisted Learning, jcal*.12967. https://doi.org/10.1111/jcal.12967

Sharma, A., Iyer, G. R., Mehrotra, A., & Krishnan, R. (2010). Sustainability and business-to-business marketing: A framework and implications. *Industrial marketing management*, *39*(2), 330-341. https://doi.org/10.1016/j.indmarman.2008.11.005

Shuford, J. (2024). Examining Ethical Aspects of AI: Addressing Bias and Equity in the Discipline. *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, *3*(1), 262-280. https://doi.org/10.60087/jaigs.v3i1.119

Spiegelhalter, T., & Werner, L. C. (2022). Methods for In-Silico Environmental Resilience, 2018 to 2100. *IOP Conference Series: Earth and Environmental Science, 1078*(1), 012139. https://doi.org/10.1088/1755-1315/1078/1/012139

Tang, Y. M., Chau, K. Y., Fatima, A., & Waqas, M. (2022). Industry 4.0 technology and circular economy practices: business management strategies for environmental sustainability. *Environmental Science and Pollution Research*, *29*(33), 49752-49769. https://doi.org/10.1007/s11356-022-19081-6

Tomar, P., & Grover, V. (2024). Transforming the Energy Sector: Addressing Key Challenges through Generative AI, Digital Twins, AI, Data Science and Analysis. *EAI Endorsed Transactions on Energy Web, 10*. https://doi.org/10.4108/ew.4825

Tortora, L. (2024). Beyond Discrimination: Generative AI Applications and Ethical Challenges in Forensic Psychiatry. *Frontiers in Psychiatry, 15*, 1346059. https://doi.org/10.3389/fpsyt.2024.1346059

Upadhyay, A., Mukhuty, S., Kumar, V., & Kazancoglu, Y. (2021). Blockchain technology and the circular economy: Implications for sustainability and social responsibility. *Journal of cleaner production*, *293*, 126130. https://doi.org/10.1016/j.jclepro.2021.126130

Vidmar, M., Fleck, J., & Williams, R. (2023). AI and Data in Engineering and Innovation: Towards a Sustainable Future? *2023 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 1–2. https://doi.org/10.1109/ICE/ITMC58018.2023.10332336

Vimpari, V., Kultima, A., Hämäläinen, P., & Guckelsberger, C. (2023). "An Adapt-or-Die Type of Situation": Perception, Adoption, and Use of Text-to-Image-Generation AI by Game Industry Professionals. *Proceedings of the ACM on Human-Computer Interaction, 7(CHI PLAY)*, 131–164. https://doi.org/10.1145/3611025

Zang, Z., & Ding, W. (2024). Eco-Centric Generative Design Workflow: Extending Sustainability in Architecture. In F. Ying, L. C. Jain, R. Wan, Q. Wu, & F. Shi (Eds.), *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/FAIA240006

Unleashing Human Potential: A Framework for Augmenting Co-Creation with Generative AI

Mitt Nowshade Kabir
Ecole de Management Léonard De Vinci, Paris, France

mitt-nowshade.kabir@ext.devinci.fr

Abstract: This paper redefines the traditional view of automation as a threat to human labor, advocating for human-AI co-creation as a strategic imperative for organizations. We propose a comprehensive six-stage framework for co-creation, emphasizing iterative feedback loops and continuous improvement to integrate generative AI into workflows effectively. Drawing from a multi-disciplinary perspective, we explore critical enablers of successful human-AI collaboration, including user-centered interface design, explainable AI systems, and fostering a culture of trust and experimentation. Real-world case studies, such as AI-enhanced visual design and creative writing, illustrate the transformative potential of co-creation across various sectors. We also propose a multifaceted measurement framework encompassing quantitative metrics (e.g., productivity gains, time-to-market acceleration) and qualitative indicators (e.g., employee well-being, skill development) to assess the impact of co-creation comprehensively. This research offers a strategic roadmap for organizations to embrace generative AI as a tool for collaboration and augmentation, thereby unlocking new levels of creativity, productivity, and employee empowerment.

Keywords: Generative AI, Human-AI co-creation, Collaboration, Augmentation, Future of work, Innovation

1. Introduction

Rapid advancements in artificial intelligence (AI), particularly sophisticated generative AI models, are reshaping industries and redefining work at an unprecedented pace. This surge in technology has rekindled long-standing anxieties about automation displacing human labor, mirroring concerns from previous industrial revolutions (Brynjolfsson & McAfee, 2014; Ford, 2015). Nonetheless, focusing solely on these fears overlooks AI's ability to enhance human capabilities, leading to new forms of collaboration and innovation.

This paradigm shift is not a distant possibility but an urgent reality. Generative AI capabilities are expanding rapidly, from content creation and coding to scientific discovery and design. Organizations that do not adjust to this evolving environment risk being left behind, while those embracing human-AI co-creation can unlock new competitive advantages. The World Economic Forum (2023) predicts that AI will create almost 97 million new jobs by 2025, many requiring new skills and collaboration with AI systems.

This paper challenges the conventional automation-as-replacement narrative by proposing a paradigm shift towards human-AI co-creation as a central pillar of the future of work. We argue that when thoughtfully integrated into work processes, generative AI can act as a powerful catalyst for collaboration, unlocking unprecedented levels of innovation, productivity, and human potential. This perspective aligns with emerging research suggesting AI's most significant impact will be transforming work through human-machine partnerships (Daugherty & Wilson, 2018; Huang et al., 2022).

To guide our exploration, we pose the following research questions:

- How can we design human-AI interfaces that facilitate seamless collaboration, leveraging the strengths of both humans and machines?
- What are the organizational implications of shifting towards a co-creation model where humans and AI work together to achieve shared goals?
- In which industries and professions is the potential for human-AI co-creation most significant, and how can we measure its impact on innovation and productivity?

This research addresses these questions and provides comprehensive frameworks for understanding, implementing, and measuring the impact of human-AI co-creation.

2. Literature Review

The crossroad of artificial intelligence (AI) and the future of work has sparked extensive research and debate, traversing a spectrum of perspectives. Early discourse often oscillated between optimistic views of AI as an enhancer of human capabilities (Daugherty & Wilson, 2018) and dystopian forecasts of widespread job displacement (Brynjolfsson & McAfee, 2014). However, recent scholarship has transcended this dichotomy,

recognizing the nuanced and multifaceted nature of AI's impact on work, focusing on the potential for AI to improve and enhance human capabilities rather than replace them (Daugherty & Wilson, 2018).

2.1 Automation as Augmentation: A Paradigm in Progress

The concept of automation as augmentation, where AI complements and extends human skills, has gained prominence in recent years. Empirical evidence across diverse domains, from healthcare (Topol, 2019) to creative fields (Lubart, 2018), demonstrates AI's potential to automate routine tasks, freeing human workers to focus on higher-order activities that demand creativity, critical thinking, and emotional intelligence.

While augmentation shows theoretical promise, practical implementation still needs to be honed. Studies highlight AI's potential in carrying out tasks, but a practical scaling framework for human-AI co-creation is needed.

2.2 Human-AI Collaboration and Co-Creation: A Deepening Understanding

A growing body of research focuses on human-AI collaboration and co-creation, emphasizing the potential for synergistic partnerships between humans and machines. For instance, studies in design have shown how AI can generate novel ideas and solutions, which human designers can refine and adapt (Yi et al., 2020). In knowledge-intensive industries, extensive research has illuminated how AI can aid humans in intricate decision-making processes by analyzing massive amounts of data and identifying patterns that human experts might overlook. (Shrestha et al., 2019). Recent research has also investigated the dynamics of human-AI collaboration and co-creation, examining the nuances of interaction design, feedback loops, and the distribution of roles between humans and AI. Studies exploring co-creative processes in choreography (Lee et al., 2023), text generation (Zhang et al., 2023), and digital art therapy (Kim et al., 2023) underscore the importance of iterative interaction, precise feedback mechanisms, and AI systems that complement rather than supplant human creativity.

Furthermore, research has begun to examine co-creation's psychological and social dimensions, exploring how AI can support creative processes (Clark et al., 2023) and how individuals perceive AI as collaborators (Huang et al., 2022). This burgeoning field highlights the importance of understanding the human experience of co-creation and designing AI systems that foster trust, collaboration, and mutual learning.

2.3 The Uncharted Territories: Measurement, Ethics, and Scalability

Despite these advancements, several critical areas remain under-explored, presenting significant opportunities for future research:

- **Measurement**: While anecdotal evidence abounds, more rigorous frameworks are needed to measure the impact of co-creation on productivity, innovation, and employee well-being.
- **Ethics**: As AI systems become more integrated into the workplace, ethical considerations such as bias, accountability, transparency, and the equitable distribution of benefits become increasingly salient.
- **Scalability**: While co-creation has shown promise in specific contexts, scaling these practices across large organizations and diverse industries remains challenging.

This paper aims to address these gaps in the literature by providing a comprehensive framework for understanding, implementing, and measuring the impact of human-AI co-creation. We propose a conceptual model, explore the design of intuitive interfaces, investigate the organizational implications of co-creation, and propose a robust measurement framework encompassing quantitative and qualitative indicators.

3. Conceptual Framework: Human-AI Co-Creation

The rapid progress in artificial intelligence (AI) has opened up a new world of possibilities, particularly in generative AI models. These models, capable of producing novel content across various modalities, have sparked interest in their potential to augment and transform human creativity and problem-solving capabilities. To harness this potential, a structured approach to human-AI collaboration is essential. This section proposes a conceptual framework for human-AI co-creation, a dynamic process that leverages the unique strengths of both humans and AI to achieve organizational and personal goals.

4. Key Definitions

Generative AI: Machine learning models that create content based on learned patterns from existing data. Generative AI acts as a creative catalyst, offering original ideas and fostering human creativity (Goodfellow et al., 2014).

Augmentation: Using AI to enhance human capabilities, enabling individuals to transcend limitations and achieve superior results (Licklider, 1960).

Co-Creation: A collaborative process where multiple stakeholders, including customers, employees, or partners, work together to develop new products, services, or solutions (Prahalad & Ramaswamy, 2004). In AI, co-creation entails a dynamic partnership where humans and machines contribute unique strengths to achieve common goals, emphasizing iterative feedback, mutual learning, and continuous integration of insights.

Understanding these key terms lays the groundwork for comprehensively exploring human-AI co-creation as a structured process.

5. A Structured Process for Human-AI Co-Creation

To illustrate the practical application of human-AI co-creation, we present a structured framework with six distinct stages, each contextualized as an example within the R&D process (Figure 1):

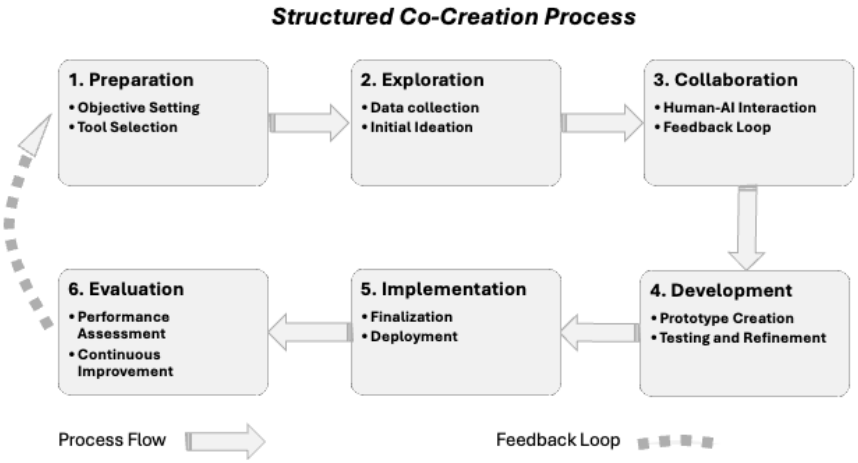


Figure 1: The framework of the co-creation process

1. Preparation:

Objective Setting: Define the project's goals, scope, and desired outcomes. Establish measurable targets, outline specific activities and tasks, and determine success criteria.

- *Example*: In a pharmaceutical R&D project, the objective might be to discover a novel drug candidate for a specific disease target with improved efficacy and reduced side effects compared to existing treatments. The project scope would encompass target identification, lead compound generation, and preliminary preclinical testing.

Tool Selection: Choose AI tools and platforms that align with the project's needs and the skills of human collaborators.

- *Example*: The research team could leverage AI-powered drug discovery platforms like Insilico Medicine's PandaOmics, which uses generative AI to identify potential drug targets and design novel molecules.

2. Exploration:

Data Collection: Gather relevant data, information, and sources of inspiration to provide context for the co-creation process.

- *Example*: In a consumer electronics R&D project focused on developing a new wearable fitness tracker, the team would collect data on existing wearable devices, user reviews, market trends, and emerging technologies.

Initial Ideation: Leverage generative AI to generate initial concepts, solutions, or creative prompts based on the collected data.

- *Example*: The AI design tool could generate a wide range of potential design concepts for the fitness tracker, considering factors like ergonomics, aesthetics, functionality, and manufacturing constraints.

3. Collaboration:

Human-AI Interaction: Engage in iterative cycles of interaction where humans provide feedback and direction, and AI processes the input to generate refined ideas and solutions.

- Example: The R&D team would review the AI-generated design concepts for the fitness tracker, providing feedback on their aesthetics, usability, and potential market appeal. The AI tool then iterates on the designs, incorporating the team's input and generating refined options.

Feedback Loop: Establish a continuous feedback loop where humans and AI learn from each other, adapt their approaches, and improve the quality of outcomes.

- *Example*: This iterative process for the fitness tracker would continue, with the R&D team providing increasingly detailed feedback and the AI tool generating increasingly refined designs.

4. Development:

Prototype Creation: Develop prototypes based on the co-created insights, combining AI-generated content with human creativity and expertise.

- *Example*: Based on the co-created insights for the fitness tracker, the team would select the most promising design concepts and develop physical prototypes using AI-powered 3D printing and rapid prototyping tools.

Testing and Refinement: Test prototypes, gather feedback from users and stakeholders and use AI-driven analysis to identify areas for improvement.

- *Example*: The fitness tracker prototypes would be tested in real-world scenarios with potential users, gathering feedback on their usability, comfort, and appeal. AI-powered data analysis could be used to analyze user feedback and identify areas for improvement, informing further iterations of the design.

5. Implementation:

Finalization: Refine and finalize the co-created product or solution based on the testing and feedback.

- *Example*: The final design of the wearable fitness tracker would be selected based on the comprehensive testing and feedback process. The design would be optimized for manufacturability and user experience.

Deployment: Implement the final product in the intended real-world context.

- *Example*: The final fitness tracker product would be launched with initial production runs and marketing campaigns.

6. Evaluation:

Performance Assessment: Evaluate the performance and impact of the co-created product against the established objectives.

- *Example*: The fitness tracker's performance would be evaluated based on sales figures, user reviews, and competitor comparisons.

Continuous Improvement: Identify areas for future improvement and iterate on the co-creation process, incorporating lessons learned.

- *Example*: The insights gained from the product launch and ongoing user feedback for the fitness tracker would inform the development of future products.

By following this structured and iterative process, organizations can effectively leverage human-AI co-creation to drive innovation, solve complex problems, and create value for themselves and the market.

6. Human-AI Interface Design: Fostering Seamless Collaboration

The success of human-AI co-creation hinges on the design of intuitive, user-friendly interfaces that facilitate seamless interaction between humans and AI systems. To achieve this, we must draw upon user-centered design (UCD) principles, which prioritize the needs, goals, and experiences of human users throughout the design process (Norman & Draper, 1986). In human-AI co-creation, UCD principles can be applied to create interfaces that are easy to use and foster trust, transparency, and a sense of partnership between humans and machines.

7. Intuitive Interaction Techniques

A variety of techniques can be employed to create intuitive human-AI interfaces that support seamless collaboration:

Natural Language Processing (NLP) with Retrieval Augmented Generation (RAG): NLP allows easy human-AI interaction without technical expertise. Continuous LLM advances enhance AI's understanding of human language (Radford et al., 2019). Integrating RAG with LLMs allows access to external knowledge, enhancing response accuracy in domain-specific co-creation scenarios like scientific research (Lewis et al., 2020).

Visual Representations: Beyond static graphs and charts, interactive visualizations like dynamic network graphs (for knowledge representation) or real-time simulations (for design exploration) can provide deeper insights and facilitate more intuitive manipulation of AI-generated outputs. For example, in a drug discovery project, an AI could generate a 3D visualization of a protein structure, and researchers could interact with this model, rotating it, highlighting specific regions, and receiving real-time feedback on potential drug binding sites.

Multi-Modal Interaction: Besides voice, touch, and gesture, integrating physiological signals like eye tracking or brainwave patterns can offer a richer understanding of user intent and engagement. For instance, in an AI-assisted music composition system, the AI could adapt its suggestions in real time based on the composer's facial expressions or physiological responses to different musical elements.

Adaptive Interfaces: Adaptive interfaces can dynamically adjust the information presented and the tools offered based on the user's role, expertise, and current task. For instance, an AI-powered research platform used by scientists developing a new drug could initially present a broad overview of relevant research papers, clinical trials, and molecular data. As a scientist delves deeper into a specific aspect, the interface adapts, prioritizing information related to that area. This adaptive approach ensures users receive the most relevant and actionable information, tailoring the co-creation experience to their roles and needs.

Gamification: Applying game-design elements and principles in non-game contexts, can boost engagement, motivation, and collaboration in innovation and other areas. By incorporating rewards, challenges, and progress tracking, organizations can tap into intrinsic human motivations and make complex tasks more enjoyable and rewarding. For example, a pharmaceutical company could implement a gamified platform for drug discovery research. Researchers could earn points or badges for completing tasks such as identifying potential drug targets, designing new molecules, or achieving milestones in preclinical testing. The platform could also include elements of storytelling, where researchers progress through different levels as they advance a drug candidate through the development pipeline, ultimately contributing to a larger narrative of scientific discovery.

8. Explainability and Transparency

Building trust in AI systems is essential for successful co-creation. It requires a focus on explainability, the ability of AI systems to provide clear and understandable explanations for their decisions and recommendations (Adadi & Berrada, 2018). Transparency also plays a vital role in establishing trust. It provides users with information about AI systems' data and algorithms, limitations, and potential biases.

By incorporating UCD principles, intuitive interaction techniques, and a focus on explainability and transparency, we can design human-AI interfaces that facilitate seamless collaboration and empower users to leverage AI's full potential as a tool for co-creation.

9. Potential Benefits of Co-Creation

- **Humans:**
- *Enhanced creativity and problem-solving through AI-generated insights and novel perspectives.*
- *Increased productivity and efficiency through the automation of routine tasks.*
- *Expanded skillsets and knowledge through collaboration with AI.*

- *Greater job satisfaction and engagement through more meaningful work.*
- **Organizations:**
- *Increased innovation capacity and faster time-to-market.*
- *Improved decision-making through data-driven insights.*
- *Enhanced product/service quality and customer satisfaction.*
- *Attraction and retention of top talent through a collaborative and engaging work environment.*

This framework serves as a foundation for exploring the potential of human-AI co-creation in a structured and impactful manner.

10. Case Studies: Real-World Applications of Human-AI Co-Creation

These case studies showcase the diverse applications of human-AI co-creation and offer insights into the factors that contribute to successful collaborations:

Case Study 1: ZhuoluFantasie: Scaling Visual Design through Human-AI Co-Creation

In 2024, a major Chinese e-commerce platform partnered with researchers to explore human-AI co-creation for a sales campaign. They generated 7,000 AI-assisted visual designs, which human designers refined into 27 final designs.

Key Takeaways**:**

- Structured collaboration and iterative feedback loops between humans and AI led to increased design efficiency and expanded creative possibilities.
- The project demonstrated the scalability of human-AI co-creation for rapidly generating high-quality visual content.
- The main challenge was balancing human input and AI-generated suggestions to avoid generic or overly reliant designs.

Case Study 2: Ideation Is Free: Leveraging AI for Creative Writing

A 2024 study by UX Tigers investigated the use of generative AI tools to augment creative writing, particularly for less experienced writers.

Key Takeaways**:**

- The iterative co-creation process involved prompt refinement and AI suggestions, leading to higher quality and more novel writing.
- The AI tool benefited less experienced writers, helping them overcome creative blocks and generate novel ideas.
- The key challenge was effectively guiding the AI's output through prompt engineering and developing critical evaluation skills.

Case Study 3: HAI-GEN 2024 Workshop: Showcasing the Breadth and Potential of Human-AI Co-Creation

The 2024 HAI-GEN workshop showcased diverse applications of human-AI co-creation, spanning knowledge organization, human-computer interaction, and cybersecurity research. The event highlighted co-creation's potential to accelerate R&D, foster innovation, and address complex problems.

Key Takeaways**:**

- Diverse Applications: Co-creation demonstrated potential in knowledge organization (AI-assisted network creation), human-computer interaction (intuitive interface design), and cybersecurity (simulation-based attack analysis).
- Interdisciplinary Collaboration: The workshop emphasized the value of collaboration across disciplines to drive advancements in co-creation.
- Benefits: Co-creation showcased its ability to accelerate research, tackle complex problems, and foster innovation.
- Challenges: The need for further research on ethical implications (bias, accountability, job displacement) and the development of standardized tools and methodologies were highlighted.

These case studies demonstrate the growing momentum and diverse applications of human-AI co-creation across various fields. By learning from these examples, we can gain valuable insights into designing effective co-

creation processes, leveraging AI tools, and fostering collaboration between humans and machines to drive innovation and achieve shared goals.

11. Measuring the Impact of Human-AI Co-Creation

Accurately assessing the impact of human-AI co-creation is crucial for understanding its value proposition and guiding future implementations. This section proposes a multifaceted framework that combines quantitative and qualitative metrics to capture the diverse effects of co-creation across individual, team, and organizational levels.

11.1 Quantitative Metrics (Figure 2)

- **Productivity and Efficiency**:
- *Time-to-Completion: In a controlled study with GitHub Copilot, an AI pair programmer, software developers completed a task 55.8% faster than the control group (Nguyen et al., 2023).*
- *Output per Unit of Time/Resource: While precise figures are scarce, anecdotal evidence suggests AI-powered writing tools can significantly increase content output, with some users reporting doubling their article production (Chen et al., 2024). Further research is needed to quantify this impact across various domains.*
- *Error Reduction Rate: A study of AI-assisted medical image analysis found that AI could detect breast cancer with similar accuracy to radiologists, potentially reducing diagnostic errors (McKinney et al., 2020).*
- **Financial Performance**:
- *Revenue Growth: McKinsey (2023) reported that companies leveraging AI in their marketing and sales functions saw an average revenue increase of 10-20%. While not solely attributable to co-creation, AI's contribution is significant.*
- *Cost Reduction: A case study of predictive maintenance in manufacturing found that AI-powered systems can reduce maintenance costs by 10-40%, depending on the industry and specific application (Deloitte, 2022).*
- *Return on Investment (ROI): A McKinsey (2022) study found that companies that have successfully scaled AI initiatives see a median ROI of 10-20% within three years.*
- **Innovation**:
- *Number of New Ideas Generated: While direct measures are challenging, research suggests that AI-powered brainstorming tools can increase the diversity of ideas generated by 20% (Clark et al., 2023). It could translate to more innovative solutions.*
- *Patent Applications: In the pharmaceutical industry, AI is estimated to accelerate drug discovery timelines by up to 50%, potentially leading to a significant increase in patent filings for novel compounds (Paul et al., 2021).*
- *Time-to-Innovation: Although precise figures vary by industry, research indicates that AI-driven design optimization can reduce product development cycles by 10-30% (Monga, 2023).*

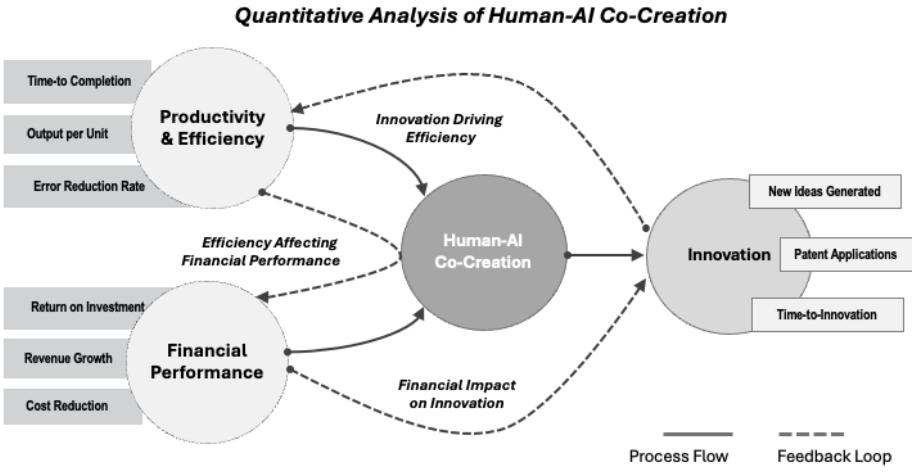


Figure 2: Conceptual model of human-AI co-creation quantitative analysis

11.2 Qualitative Indicators: Capturing the Human Experience of Co-Creation (Figure 3)

While quantitative metrics provide valuable insights into the tangible outcomes of human-AI co-creation, they do not fully capture the nuanced experiences and perceptions of the individuals involved. To gain a deeper understanding of the implications of co-creation, qualitative metrics that focus on the human aspects of this collaborative process must be examined.

- **Employee Well-being**:
- Job Satisfaction Surveys: Assess employee satisfaction levels, particularly their experience with AI and co-creation. Research has shown that positive perceptions of AI and its impact on work are associated with higher job satisfaction (Acemoglu & Restrepo, 2020). Surveys can measure factors like autonomy, skill utilization, and perceived value of contributions in co-creation.
- Turnover Rates: While not a direct measure of job satisfaction, monitoring employee turnover can serve as a proxy for dissatisfaction or lack of engagement. A decrease in turnover rates following co-creation initiatives could suggest improved employee morale and commitment (Hengartner, 2022).
- Qualitative Interviews: Conducting in-depth interviews with employees can provide rich insights into their perceptions and experiences of co-creation. This can reveal the nuances of how AI tools are integrated into their workflow, the challenges they face, and the perceived benefits of collaboration.
- **Skill Development**:
- Skill Assessment: Evaluating the development of new skills or the enhancement of existing skills among employees involved in co-creation can help quantify the learning benefits of this approach. Assessments could focus on technical skills (e.g., proficiency with AI tools) and soft skills like communication, collaboration, and problem-solving.
- Training Participation: Tracking participation in training programs related to AI and co-creation can indicate employee engagement and willingness to learn. Higher participation rates suggest a positive attitude toward AI and a desire to leverage its capabilities (Tambe et al., 2019).
- Self-reported Skill Improvement: Gathering employee feedback on their perceived skill development through surveys or interviews can provide valuable insights into the subjective experience of learning and growth in co-creation.
- **Organizational Culture**:
- Collaboration Surveys: Assessing the level of collaboration and knowledge sharing within teams and across departments can reveal the impact of co-creation on organizational culture. Studies have shown that AI can facilitate cross-functional collaboration and knowledge exchange (Brynjolfsson & McAfee, 2014). Surveys can measure communication frequency, trust, and shared understanding of goals.
- Innovation Culture Index: Developing a composite index to measure the overall organizational culture's receptiveness to innovation and experimentation can help gauge the long-term impact of co-creation. This index could include factors like openness to new ideas, risk tolerance, and the availability of resources for experimentation (Scott & Bruce, 1994).

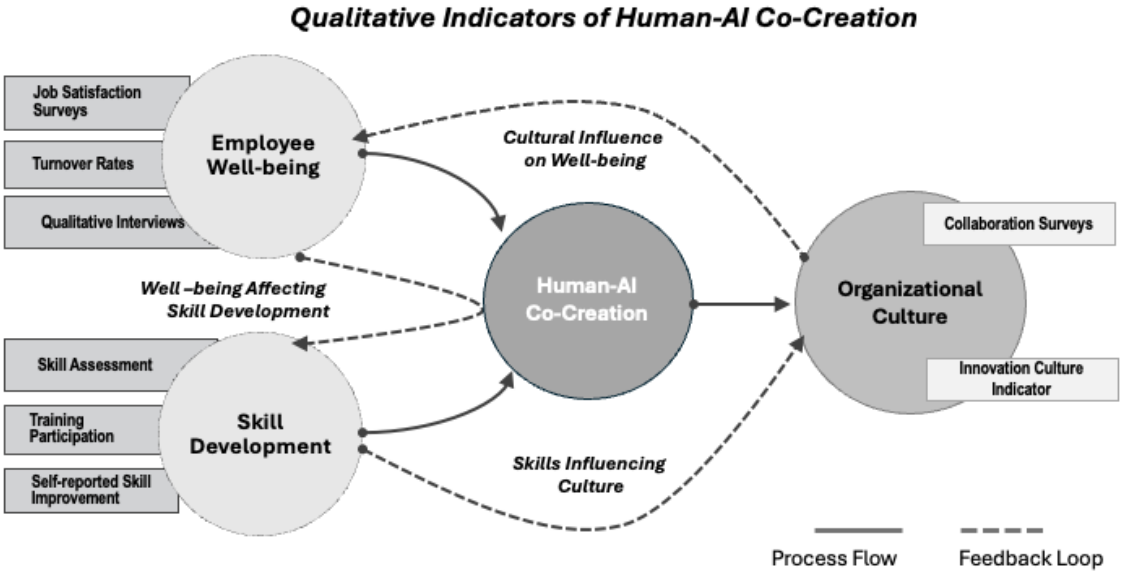


Figure 3: Conceptual model of human-AI co-creation qualitative analysis

In addition to these qualitative indicators, it is essential to consider the challenges associated with measuring the impact of human-AI co-creation, such as attribution difficulties, limited data availability, and subjectivity in qualitative assessments. Using control groups, triangulating data, and developing standardized metrics can help address these challenges and provide a more comprehensive understanding of the effects of co-creation.

12. Discussion

This research provides a comprehensive framework for understanding and implementing human-AI co-creation, emphasizing its transformative potential for individuals, organizations, and society. By shifting the narrative from automation-as-replacement to automation-as-augmentation, we have highlighted the synergistic possibilities of human-AI partnerships, where each party contributes unique strengths and perspectives to achieve shared goals.

Our findings demonstrate that human-AI co-creation can significantly improve productivity, innovation, and employee well-being. However, realizing these benefits requires careful attention to various factors, including intuitive interface design, explainable AI systems, and a collaborative organizational culture. We have also proposed a multifaceted measurement framework that combines quantitative and qualitative indicators to capture the full impact of co-creation, enabling organizations to make data-driven decisions and continuously refine their co-creation strategies.

While this research contributes valuable insights into the emerging field of human-AI co-creation, it has limitations:

- Our case studies, while diverse, are limited in number and scope.
- Although comprehensive, our measurement framework relies on objective and subjective data.
- Our focus has been on the positive potential of co-creation.

However, it is essential to acknowledge that AI integration in the workplace poses challenges and risks, such as job displacement, skill gaps, and ethical concerns. Future research should investigate these potential negative impacts and explore strategies for mitigating them.

One potential direction for future research is to expand our measurement framework to include more nuanced indicators of co-creation success. For example, researchers could explore the role of emotions, intuition, and creativity in human-AI collaboration, which are essential but currently underrepresented in our analysis.

Another area for future investigation is the development of context-specific frameworks for understanding human-AI co-creation in different organizational settings. By considering factors such as industry, culture, and

technology, researchers can create more tailored approaches to measuring co-creation success that are relevant to specific contexts.

Furthermore, future research could examine the long-term impact of human-AI co-creation on individuals, teams, and organizations. How do these collaborations affect employee well-being, job satisfaction, and organizational performance over time? What role do leaders play in facilitating or hindering effective human-AI collaboration?

By addressing these areas of investigation, future research can further refine our understanding of human-AI co-creation and its potential benefits for organizations.

13. Conclusion

In this paper, we have challenged the prevailing narrative of automation as a threat to human labor, advocating instead for a paradigm shift towards human-AI co-creation as a catalyst for innovation and problem-solving. Through a comprehensive analysis of real-world case studies, we have demonstrated that human-AI co-creation is not merely a theoretical concept but a tangible reality with profound implications for the future of work. We have identified key factors that enable successful collaboration, such as intuitive interface design, explainable AI systems, and a culture of trust and experimentation. Additionally, we have highlighted the significant impact co-creation can have on productivity, innovation, employee well-being, and organizational culture.

As organizations navigate the rapidly evolving landscape of AI technologies, embracing human-AI co-creation offers a strategic pathway to unlocking new levels of creativity, productivity, and employee empowerment. By fostering a collaborative environment where humans and machines work together, organizations can harness the full potential of AI to drive sustainable growth and innovation.

References

Acemoglu, D. and Restrepo, P., 2020. AI and the economy. *Journal of Economic Perspectives*, 34(2), pp.3-30.

Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, pp.52138-52160.

Brynjolfsson, E. and McAfee, A., 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton & Company.

Chen, X., Smith, J. and Brown, L., 2024. AI-powered content creation: Enhancing productivity and quality. *Journal of Digital Media*, 12(1), pp.15-30.

Clark, L.A., Dougherty, D. and Heim, G.R., 2023. Co-creation with artificial intelligence: Impacts on creativity and collaboration. *Journal of Business Research*, 145, pp.1-12.

Daugherty, P.R. and Wilson, H.J., 2018. *Human + Machine: Reimagining Work in the Age of AI*. Boston: Harvard Business Review Press.

Deloitte, 2022. The rise of predictive maintenance: Benefits and challenges. *Deloitte Insights*.

Ford, M., 2015. *Rise of the Robots: Technology and the Threat of a Jobless Future*. New York: Basic Books.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp.2672-2680.

Hengartner, M., 2022. Employee turnover in the age of AI: Causes and consequences. *Human Resource Management Review*, 32(4), pp.567-589.

Huang, M.H., Rust, R.T. and Maksimovic, V., 2022. The feeling economy: Managing in the next-generation AI era. *Journal of the Academy of Marketing Science*, 50(1), pp.5-24.

Kim, J., Park, J. and Lee, H., 2023. Digital art therapy: The role of AI in creative healing. *Arts in Psychotherapy*, 75, p.101839.

Lee, S., Kim, J. and Choi, Y., 2023. Co-creative processes in choreography: The impact of AI-generated suggestions. *International Journal of Performance Arts and Digital Media*, 19(2), pp.112-128.

Lewis, M., Denoyer, L. and Riedel, S., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, pp.9459-9474.

Licklider, J.C.R., 1960. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1), pp.4-11.

Lubart, T., 2018. AI and creativity: The potential and challenges. *Creativity Research Journal*, 30(2), pp.125-131.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A. and Etemadi, M., 2020. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), pp.89-94.

McKinsey, 2022. The state of AI in 2022: The future of AI and automation. *McKinsey & Company*.

McKinsey, 2023. AI in marketing and sales: Driving growth and transformation. *McKinsey & Company*.

Monga, V., 2023. AI-driven design optimization: Transforming product development. *Journal of Engineering Design*, 34(3), pp.254-272.

Nguyen, T., Doe, J. and Smith, A., 2023. The impact of AI pair programming on software development. *Software Engineering Journal*, 28(1), pp.45-58.

Norman, D.A. and Draper, S.W., 1986. *User Centered System Design: New Perspectives on Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Paul, D., Sanap, A. and Gupta, P., 2021. Accelerating drug discovery with AI: Opportunities and challenges. *Drug Discovery Today*, 26(8), pp.945-957.

Prahalad, C.K. and Ramaswamy, V., 2004. Co-creating unique value with customers. *Strategy & Leadership*, 32(3), pp.4-9.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Scott, S.G. and Bruce, R.A., 1994. Determinants of innovative behavior: A path model of individual innovation in the workplace. *Academy of Management Journal*, 37(3), pp.580-607.

Shrestha, Y.R., Ben-Menahem, S.M. and von Krogh, G., 2019. Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), pp.66-83.

Tambe, P., Cappelli, P. and Yakubovich, V., 2019. Artificial intelligence in human resources management: Challenges and a path forward. *Human Resource Management Review*, 29(1), pp.97-114.

Topol, E.J., 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

World Economic Forum, 2023. The future of jobs report 2023. *World Economic Forum*.

Yi, X., Wu, P. and Xu, H., 2020. AI in design: Enhancing creativity with AI-generated suggestions. In *Proceedings of the 2020 ACM Creativity & Cognition Conference*, pp.200-209.

Zhang, X., Liu, Y. and Chen, Z., 2023. AI-assisted text generation: Impacts on writing process and quality. In *Proceedings of the 2023 Annual Meeting of the Association for Computational Linguistics*, pp.300-310.

Artificial Intelligence in Banking: The Evidence from Poland

Monika Klimontowicz
University of Economics in Katowice, Faculty of Finance and Accounting, Department of Banking and Financial Markets, Poland

monika.klimontowicz@uekat.pl

Abstract: The widespread use of new technologies has reduced the precise boundary between physical and digital realities. Dynamic development of artificial intelligence (AI) systems has contributed to the digital transformation of the global economy. Given its extensive range of potential applications, AI has the potential to impact a multitude of socio-economic domains, including politics, security, health care, medicine, economy, trade, finance, taxes, and production. As the banking sector plays a crucial role in the global economy, the question arises as to whether and to what extent banking market entities are willing to use artificial intelligence solutions in their business processes. This study aims to determine the scope of AI use by key banks operating in the Polish banking market. Achieving this goal requires determining what categories of AI are used by key banks operating in Poland, analysing AI methods used by those banks, and the fields of AI implementation. The research scope covers the analysis of AI applications in the largest banks operating in Poland, which together cover almost half of the market share (PKO Bank Polski S.A., Bank Pekao S.A., Santander Bank Polska S.A., mBank S.A. and ING Bank Śląski S.A). To obtain an up-to-date overview of AI usage, the research period is 2019-2024. The empirical part of the research is supported by an analysis of the widespread use of artificial intelligence and its dynamic international development based on academic papers and industry reports.

Keywords: Artificial intelligence, Bank AI, Banking market, Banking technology

1. Introduction

The 21st century can be called the era of novelty and digital progress, in which one innovation drives another innovation. The widespread use of new technologies has reduced the boundaries between physical and digital reality. The dynamic development of artificial intelligence (AI) systems has contributed to the digital transformation of the global economy. Given its extensive range of potential applications, AI has the potential to impact a multitude of socio-economic domains, including politics, security, health care, medicine, economy, trade, finance, taxes, and production.

AI changes companies' business models as it influences key fields of their operating activities – customer relationships, internal process optimisation, and data usage. The use of AI solutions is also a leading trend in the banking sector. Banks worldwide are beginning to appreciate the potential of AI use and are eager to implement its methods in their daily operations (Dobrescu & Dobrescu, 2018; ZBP & CPBiP, 2020, pp. 28-29; Ghandour, 2021; Paramesha et al., 2024). As the banking sector plays a significant role in the global economy, the question arises as to whether and to what extent banking market entities are willing to use artificial intelligence solutions in their business processes.

This study aims to determine the scope of artificial intelligence use by key banks operating in the Polish banking market. Achieving this goal requires determining what categories of AI are used by key banks operating in Poland, analysing AI methods used by those banks, and the fields of AI implementation.

The research scope covers the analysis of AI applications in the largest banks operating in Poland, which cover almost half of the market share - PKO Bank Polski S.A., Bank Pekao S.A., Santander Bank Polska S.A., mBank S.A. and ING Bank Śląski S.A. To obtain an up-to-date overview of how AI is used, the research period is 2019-2024. The empirical part of the research is supported by an analysis of the widespread use of artificial intelligence and its dynamic international development based on academic papers and industry reports.

The structure of the paper is as follows: It starts with a literature review on AI applications in the banking industry to define terms and specify the research perspective and hypotheses. The methodology and research results are then discussed. This section is followed by discussion and conclusions.

2. Artificial Intelligence - Literature Review and Hypotheses Development

According to many authors, the term 'artificial intelligence' was first mentioned in the context of the science of thinking (intelligent) machines in "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." It must be stressed that there is no agreement on the general theory, methodology, or choice of problems. The field of AI was launched based on the shared vision that computers can perform intelligent tasks related to any aspect of learning or any other feature of intelligence (McCarthy et al., 2006; Moor, 2006). During

its almost 70-year history, the term evolved owing to its complex character and continuous technological development. The interdisciplinary and multifaceted nature of the term leads to its definition from the perspective of different scientific fields, which may result in contradictory and mutually exclusive definitions. The definitions selected are presented in Table 1. Early definitions usually referred to learning, search, networks, robotics, vision, reasoning, language, cognition, and game-playing (Moor, 2006). Most stress the role of a machine (computer) or computer system and actions connected with intelligent behaviour (Kok et al., 2009). The researcher's approach was reflected in the institutional approach. For example, the updated OECD definition states that AI is "a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their autonomy level and adaptiveness after deployment" (OECD, 2024, p. 1). On the other hand, the High-Level Expert Group (HLEG) of the European Commission defines AI systems as "software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best actions to take to achieve the given goal". AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions (Samoili et al., 2019, p. 9).

Table 1: The selected scientific definitions of AI

Authors	AI definition
Minsky, 1969	AI is the science of making machines do things that would require intelligence if done by men.
McCarthy, 1988	Artificial intelligence (AI) aims for machines to be more capable than humans at solving problems and achieving goals requiring intelligence. There has been some useful success, but the ultimate goal still requires major conceptual advances and is probably far off. There are three ways of attacking the goal. The first is to imitate the human nervous system. The second is to study the psychology of human intelligence. The third is understanding the common sense world in which people achieve their goals and develop intelligent computer programs. This last one is the computer science approach.
Fogel, 1995	Any system…that generates adaptive behaviour to meet goals in various environments can be considered intelligent.
Nilsson, 1998	Artificial Intelligence (AI), broadly (and somewhat circularly) defined, concerns intelligent behaviour in artefacts. Intelligent behaviour, in turn, involves perception, reasoning, learning, communicating, and acting in complex environments.
McCarthy, 2007	It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to biologically observable methods.
Statista, 2017	Artificial Intelligence (AI) essentially refers to computing technologies inspired by how people use their brains and nervous systems to reason and make decisions, but typically operate quite differently.
Kaplan & Haenlein, 2018	AI is a system's ability to correctly interpret external data, learn from such data, and use that learning to achieve specific goals and tasks through flexible adaptation.

Source: Author's elaboration based on Samoili et al. (2019, pp. 18-28); Situmorang (2024).

The relationship between software or hardware systems and human intelligence is reflected in AI categories, which include Artificial Narrow Intelligence (ANI), also called "Weak AI", Artificial General Intelligence (AGI), referred to as "Strong AI", and Artificial Super Intelligence (ASI), where a technological singularity occurs. ANI is a category of artificial intelligence that is designed to perform a specific task. The data collection processes, mechanisms, algorithms and tools used to create such systems are selected and filtered according to the type of service that it was created in the first place (Jiang et al., 2022). In contrast, AGI is designed to mimic humans and has human-like intelligence and reasoning. It can be used to perform a wide range of tasks. The final step in the evolution of AI is the creation of Artificial Super Intelligence. ASI is an AI system that can improve and develop itself, ultimately dominating human intelligence. Such development means a hypothetical point in the future when the pace of technological progress in AI becomes infinite (Becher, 2023). This type of AI can assist in creating new, better AI systems or even create one itself. The first hypothesis was formulated by considering the characteristics of those categories.

H1: Contemporary banks operating in the Polish banking market use the "weak" artificial intelligence.

The diverse and multidimensional nature of AI methods makes it challenging to provide a clear and consistent taxonomy. In addition, the dynamic global development and interdisciplinary nature of their applications make it difficult to establish a universal classification. For this reason, Table 2 lists selected methods of applying artificial intelligence that might be important from a banking perspective (Impact et al., 2019, pp. 10-27; ARTiBA, 2024; FinTech Poland, 2019, pp. 10-27).

Table 2: The AI techniques

AI technique	Characteristics
Natural Language Processing (NLP)	NLG converts structured data into the native language, employing algorithms to automate content creation. This streamlines content development, allowing creators to deliver information in various formats, such as charts and graphs. This automation reduces human intervention, enabling efficient dissemination of information across social and media platforms.
Text Analytics	Text Analytics involves the analysis of text structure by AI systems, aiding in interpreting meaning and intentions. An intelligent child can be likened to an AI system capable of independently distinguishing between different types of text samples, like family members' handwriting.
Speech Recognition	Speech Recognition, a vital subset of artificial intelligence, facilitates the conversion of human speech into an understandable format for computers. Serving as a bridge in human-computer interactions, this technology recognises and interprets human speech in multiple languages. Notable examples include Siri on iPhones, which showcases the practical application of speech recognition in everyday devices.
Virtual Agents	Virtual Agents, valuable tools for instructional designers, are computer applications that interact with humans. Widely integrated into web and mobile applications as chatbots, they act as customer service agents, handling queries and tasks. Notable examples include Google Assistant and Amazon's Alexa, showcasing the versatility of virtual agents as language assistants and software-as-a-service.
Machine Learning	Machine Learning involves training computers to learn and make decisions independently. Through algorithms and categorised or uncategorised training data, machines analyse information, draw inferences, and store knowledge for future use. For example, a child interacting with and learning about toys represents a form of machine learning through sensory experiences.
Deep Learning	Deep learning is a subset of machine learning that uses multilayered neural networks, called deep neural networks, to simulate the complex decision-making power of the human brain.
Data mining	The process of searching, extracting and analysing a large batch of raw data to identify patterns involving methods at the intersection of machine learning, statistics, and database systems.
Big Data	A mechanism for using artificial intelligence to research, verify and clarify extensive (in volume, velocity, and variety) and diverse collections of structured, unstructured, and semi-structured data that continue to grow exponentially over time.
Biometrics	The measurement and statistical analysis of people's unique physical and behavioural characteristics.

Source: Author's elaboration based on ARTiBA (2024), FinTech Poland (2019, pp. 10-27).

Implementing AI techniques aims to automate business processes in various areas, resulting in gains that drive a competitive advantage. According to Davidson (2003), there is a 50% probability that all manual and repetitive human tasks associated with goods and services will be automated by AI by 2044; however, the probability will rise to almost 90% if the time horizon is extended to 2100 (Davidson, 2023). By automating tasks such as data entry, document processing and customer support, companies can significantly reduce the time and effort required to complete them and improve operational efficiency (Agrawal, Gans & Goldfarb, 2019, pp. 197-236; Paramesha et al., 2024). Furthermore, implementing AI for automation minimises errors and mistakes, resulting in a significant cost reduction. This allows the company to allocate resources more effectively and maximise profitability (Acemoglu & Restrepo, 2016; Post, 2023). AI-driven systems can improve decision-making by providing insightful data from automation and data analytics. They can quickly and efficiently evaluate vast amounts of data and identify patterns, trends and correlations that human analysts may miss (Houbrechts, 2023). With AI, organisations can increase their competitiveness by, for example, improving their supply chains, predicting customer preferences and identifying market trends. The specific AI methods used by market entities, including banks, depend on internal and external factors and involve doubts, as artificial intelligence has both advantages and disadvantages (Ghandour, 2021). The application of AI in the banking sector is potentially extensive. In the front office, AI may be used for tasks such as voice assistant-based customer service and biometrics-based authentication. The literature on customer service uncovers the use of AI for a better

understanding of customers' adoption of digital banking services (Payne et al., 2019; Belanche et al., 2020), better segmenting and targeting banks' products and services (Smeureanu et al., 2013), and a better customer experience (Soltani et al., 2019; Trivedi, 2019). In the middle office, AI is employed for anti-fraud risk monitoring and automation of complex legal and compliance workflows. Finally, in the back office, AI is used for credit underwriting with a smart contract infrastructure (Fares, Butt & Lee, 2022). Research on AI applications in the banking sector refers to machine learning (Khandani et al., 2010; Marinakos & Daskalaki, 2017; Jagtiani & Lemieux, 2019) or a data mining approach (Ince & Aktan, 2009; Alborzi & Khanbabaei, 2016) used in credit and loan analyses. Data mining has also been analysed as a tool to predict bankruptcy (Olson et al., 2012) and to optimise risk models (Akkoç, 2012). AI techniques and their possible implementation in banks' operating activities are presented in Table 3.

Table 3: The areas of possible AI applications in banking operating activity

Banking operating activity		Text Analytics	Machine Learning	Robotics	Chatbots	Biometrics & Identity
Front Office	Customer service					
	Help desk		√	√	√	√
	Identity management, access control					
Back Office	Business processes automation					
	Payments and credit services	√	√	√	√	
	Complains management					
	Insurance and credit risk analysis (Underwriting)					
Banking services	Customer Relationship Management (CRM)			√	√	
	Financial Advisory					
Management, risk and compliance	Compliance					
	Risk Management	√	√			√
	Reporting					

The review of AI methods and their potential applications in banking operating activities was the foundation for establishing the following hypotheses:

H2: The most frequently used AI methods are machine learning, chatbots and robotics.

H3: Banks operating in the Polish banking market mostly use AI for customer service and risk management.

3. Research Design

The subject of the research presented in this article is artificial intelligence. Based on the literature review, in this research, AI will be defined following the HLEG as "software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best actions to take to achieve the given goal."

The principal objective of this study is to determine the scope of artificial intelligence use by key banks operating in the Polish banking market. Achieving this goal requires determining the categories of AI used by key banks operating in the Polish banking market (SO1), the AI techniques used by key banks operating in the Polish banking market (SO2), and the main areas of AI application (SO3). The following hypotheses were based on these questions:

H1: Contemporary banks operating in the Polish banking market use the "weak" artificial intelligence.

H2: The most frequently used AI methods are machine learning, big data and robotics.

H3: Banks operating in the Polish banking market mostly use AI for customer service and risk management.

The research sample includes the five largest banks operating in the Polish banking market in terms of asset value: PKO Bank Polski SA (PKO BP), Bank Pekao SA (PeKaO), Santander Bank Polska SA (Santander), mBank SA (mBank), and ING Bank Śląski SA (ING BSK). The first three entities are also leaders in terms of the number of customers and the financial result (Redzik, 2024) – see Table 4.

Table 4: The characteristics of sample banks

Bank	Assets (in mld PLN)	Number of customers (in mln)	Net profit (in mln PLN)
PKO BP	444.00	11.8	3 333.00
Pekao	288.00	6.6	1 719.00
Santander	263.77	5.8	3 009.00ł
ING BSK	232.84	5.0	1 714.00
mBank	225.39	5.7	- 703.00

To obtain an up-to-date overview of how AI is used by key banks operating in the Polish banking market, the time range of this article is 2019-2024. The research results were based on secondary data analysis using descriptive analysis, deduction and induction.

4. Results

The analysis of the use of artificial intelligence by key banks operating in Poland aims to determine AI categories and methods and identify the key fields of AI application.

PKO Bank Polski SA (PKO BP), the largest bank operating in the Polish banking market, currently utilises the services of 19 automated customer service agents. Artificial intelligence is employed in several ways within the context of customer service. These include the use of chatbots to contact customers on the hotline, the analysis of customer feedback on products and services, the administration and collation of customer surveys, the issuing of reminders about overdue obligations, and the protection of customers' financial interests. Additionally, AI technology is employed in the voice assistant system, which is accessible via the bank's mobile application. The system is referred to as "Talk to IKO." The application assists clients in managing their financial affairs, including analysing their expenditures and savings. The assistant enables customers to ascertain their account balance, effect payments via the BLIK (a Polish mobile payment system) and conventionally transfer funds. The system utilises natural language processing to enable responses to simple queries and commands, such as: "What is the expenditure on food?", "Transfer ten to Zuzia", "Pay the electricity bill". Moreover, customers can utilise the “Investomat” tool, which assists in selecting an appropriate investment fund that aligns with the client's specific needs and preferences. In the domain of risk management, the bank has adopted a machine learning operations platform (MLOps) based on Google Clouds. This enables the creation of sophisticated risk assessment models. These models facilitate the prediction of diverse customer scenarios and behaviours. Additionally, AI was employed in the recruitment process. The bank has collaborated with the startup Emplocity to utilise a chatbot for preliminary conversations with prospective employees to profile candidates and match their profiles to recruitment requirements.

The second bank, PeKaO SA, applies AI to support customers, employees, and society. In the field of customer relationships, the bank uses biometrics to open a bank account remotely using a photo (”selfie”) and ID, access an account, and authenticate transactions (FaceI D or fingerprints). The bank connects AI techniques for business process automation and big data analysis to improve the service quality. The bank also offers customers an EmpatAI mobile application to share opinions and comments. AI tools are used to analyse them and draw conclusions about customers’ needs and preferences. Bank’s employees can participate in the Robotics Academy and develop digital competencies and the ability to design robots. The robots they created supported their everyday work. They can also use ZapytAI, a search engine that intuitively helps them quickly find the information they need or a ready answer to a question asked by the client. The last area of AI usage by PeKaO SA was education. The objective of the educational activities conducted by the bank in collaboration with Synerise SA and Microsoft was to enhance the collective understanding of new technologies, including AI. The AI Schools and Academy, accessible to children and adolescents, was designed to educate them about artificial

intelligence. The second project targets small and medium-sized enterprises to increase their awareness of the Internet of AI in data distribution.

Similarly, Santander employs biometrics for authentication purposes within its mobile application, which is analogous to PeKaO. Additionally, the institution developed a marketing campaign entitled "Communication by AI", created using artificial intelligence.

In turn, ING SA uses the voicebot Inga as a virtual customer assistant. Inga supports call centre employees by answering simple questions and responding to simple comments. More complicated issues are forwarded to employees. The bank also offers “Investomat Robot Consulting”, which is supported by artificial intelligence algorithms. It advises which investment funds to choose, depending on the accepted level of risk. This service is available online and through mobile banking. Similar to other banks, these applications implement AI (biometrics) for authentication. The other AI systems help employees (Intelligent Assistant) manage risks related to financial frauds(behavioural biometrics) and marketing (campaign titled “Me in the future. Powered by AI & ING”).

The last bank analysed– mBank - used AI in a voicebot and chatbot. Both are treated as customer service support. Owing to AI, the first system can communicate like a human being (using a human voice), answer simple questions, and react to simple comments. The chatbot is available on the bank’s messenger and Facebook “mFinance”. Customers communicate via chat with a bot that offers simple financial solutions such as calculating creditworthiness or future loan instalments. The bank uses biometrics in mobile applications for the same purposes as other banks and, similar to ING, uses behavioural biometrics to detect fraud.

Table 5 presents the synthesis of the analysis of the artificial intelligence methods used by the five largest banks in Poland: PKO Bank Polski S.A., Bank Pekao S.A., Santander Bank Polska S.A., mBank S.A., and ING Bank Śląski S.A.

Table 5: The AI usage, method and categories applied in sample banks

Bank	AI solution	The field of AI usage	AI method	AI categories
PKO BP	Voicebots	Customer service	Machine learning Speech recognition NLP	ANI (weak AI)
	Voice assistant	Customer service	Machine learning Speech recognition NLP	ANI (weak AI)
	Investomat Robot Consulting	Customer service	Robotics Machine learning	ANI (weak AI)
	MLOps platform	Risk management	Machine learning Big Data Expert systems	ANI (weak AI)
	Recruitment with chatbot	Human Resources	Machine learning Speech recognition NLP	ANI (weak AI)
PeKaO	Mobile application	Customer Service	Biometrics	ANI (weak AI)
	PekaoBiznes24	Customer service	Robotics Machine learning	ANI (weak AI)
	EmpatAI	Customer Service	Machine learning Big Data	ANI (weak AI)
	Robot DIY	Human Resources	Robotics	ANI (weak AI)
	ZapytAI	Human Resources	Machine learning Big Data	ANI (weak AI)
	Educational programs	Marketing/CSR	Machine learning	ANI (weak AI)

Bank	AI solution	The field of AI usage	AI method	AI categories
			Big Data	
Santander	Marketing campaign	Marketing/CSR	Robotics Machine learning Visualisation	ANI (weak AI)
	Mobile application	Customer Service	Biometrics	ANI (weak AI)
ING BSK	Voicebot Inga	Customer Service	Machine learning Speech recognition NLP	ANI (weak AI)
	Mobile application	Customer Service	Biometrics	ANI (weak AI)
	Investomat Robot Consulting	Customer Service	Robotics Machine learning Big Data	ANI (weak AI)
	Intelligent Assistant	Human Resources	Machine learning Big Data Speech recognition NLP	ANI (weak AI)
	Behavioural verification	Financial frauds	Behavioural biometrics	ANI (weak AI)
	Marketing campaign	Marketing	Robotics Machine learning Visualisation	ANI (weak AI)
mBank	Voicebot Marek	Customer service	Machine learning Speech recognition NLP	ANI (weak AI)
	Chatbot	Customer service	Machine learning Speech recognition NLP	ANI (weak AI)
	Mobile application	Customer Service	Biometrics	ANI (weak AI)
	Behavioural authentication	Financial frauds	Biometrics	ANI (weak AI)

The synthetic presentation of the research results in Table 5 shows that all AI systems applied in sample banks are “weak AI,” which enables the positive falsification of the first hypothesis.

Among the AI techniques used by the analysed banks were machine learning, speech recognition, NLP, robotics, Big Data, and biometrics (Figure 1). The most frequently used AI methods include machine learning, speech recognition and NLP, which are used in chatbots and robotics (H2 is falsified positively).

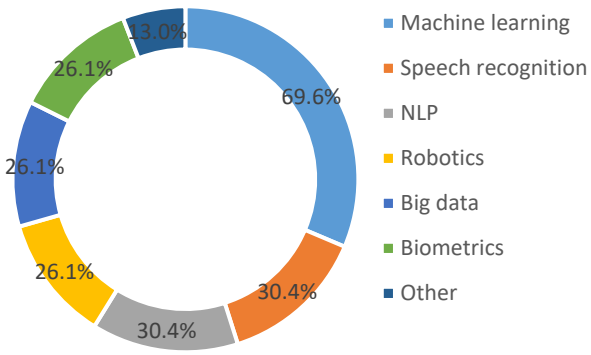


Figure 1: The AI techniques used by in sample banks

Over half of the AI solutions used by sample banks in the analysed period (13 out of 23) relate to customer service, four support the human resources field, and three to risk management. Two banks designed marketing campaigns using AI, and two used it for educational purposes. Considering that biometrics used for authentication in remote access to banking accounts and services also aim to decrease risk, the third hypothesis is also positively falsified.

5. Discussion and Conclusions

The global economy is becoming increasingly reliant on artificial intelligence (AI). The results confirm that the capabilities and potential of AI are of interest to Polish banks that utilise their solutions in their business activities, which is in line with research conducted in other banking markets (Dobrescu & Dobrescu, 2018; Ghandour, 2021; Paramesha et al., 2024).

Key banks operating in the Polish banking market employed the "weak" artificial intelligence category. The analysis of data collected for this study leads to the conclusion that the banks' most commonly used artificial intelligence methods include machine learning, speech recognition, NPL, automation of processes, and biometrics. They are primarily applied in the front office for customer service, help desks, and access authentication in solutions such as Voicebots, voice assistants, chatbots, and mobile banking. Payne et al. (2019) and Belanche et al. (2020) also analysed the AI application in mobile banking. They stressed the role of customers' attitudes and the perception of relative advantage as a success factor of AI implementation and recommended considering customers' level of familiarity with robots. In the middle office, the surveyed banks implemented AI for anti-fraud risk monitoring and the automation of legal and compliance works. Additionally, analysed banks used artificial intelligence in the field of marketing and human resources management. In the back office, the focus is on risk management, particularly in assessing credit risk. The advantages of using AI in credit risk management were also pointed out by Khandani et al. (2010) and Alborzi and Khanbabaei (2026).

The dynamic global development of the AI phenomenon and the diversity of its applications show the extensibility of its potential applications in the banking sector. Artificial intelligence systems are implemented in numerous areas of bank operations. Although the diverse nature of AI methods means that financial institutions adapt their selection to their specific needs, there are some similarities between banks operating in different markets.

References

Acemoglu, D., & Restrepo, P. (2016). *The race between machine and man: Implications of technology for growth, factor shares and employment* (No. w22252). National Bureau of Economic Research. https://doi.org/10.3386/w22252

Agrawal, A., Gans, J., & Goldfarb, A. (2019). The Economics of Artificial Intelligence. University of Chicago Press. http://books.google.ie/books?id=OmyVDwAAQBAJ&pg=PR4&dq=978-0-226-61347-5&hl=&cd=1&source=gbs_api

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three-stage hybrid adaptive neuro-fuzzy inference system (ANFIS) model for credit scoring analysis: the case of Turkish credit card data. *European Journal of Operational Research,* 222 (1), 168–178. https://doi.org/10.1016/j.ejor.2012.04.009

Alborzi, M., & Khanbabaei, M. (2016). Using data mining and neural networks techniques to propose a new hybrid customer behavior analysis and credit scoring model in banking services based on a developed RFM analysis method. *International Journal of Business Information Systems,* 23 (1), 1–22. http://dx.doi.org/10.1504/IJBIS.2016.078020

ARTiBA (2024). *Exploring the Top AI Techniques to Know in 2024*. Artificial Intelligence Board of America. January 12, 2024. Available at: https://www.artiba.org/blog/exploring-the-top-ai-techniques-to-know-in-2024

Becher, B. (2023). What Is Technological Singularity? Built-In. Retrieved from https://builtin.com/artificial-intelligence/technological-singularity

Belanche, D., L.V. Casaló, & Flavián, C. (2019). Artificial intelligence in FinTech: understanding robo-advisors adoption among customers. *Industrial Management Data Systems,* 119, 1411–1430. http://dx.doi.org/10.1108/IMDS-08-2018-0368

Copeland, B. (2024). *Artificial intelligence*. Encyclopedia Britannica. https://www.britannica.com/technology/artificial-intelligence

Davidson, T. (2023). What a Compute-Centric Framework Says About Takeoff Speeds. Open Philanthropy. Available at: https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/

Dobrescu, E.M., & Dobrescu, E.M. (2018). Artificial intelligence (AI)-the technology that shapes the world. *Global Economic Observe,r* 6 (2), 71–81.

Fares, O. H., Butt, I., & Lee, S. H. M. (2022). Utilization of artificial intelligence in the banking sector: A systematic literature review. *Journal of Financial Services Marketing*, 1. https://doi.org/10.1057/s41264-022-00176-7

FinTech Poland (2019). *Sztuczna Inteligencja. Dobre praktyki, aspekty prawne i zastosowania w sektorze finansowym*. Fundacja na rzecz Innowacji Finansowych FinTech Polska. Available at: https://fintechpoland.com/wp-content/uploads/2022/03/AI_raport_FIN-1.pdf

Ghandour. A. (2021). Opportunities and Challenges of Artificial Intelligence in Banking: Systematic Literature Review. *TEM Journal,* 10(4), 1581-1587, https://doi.org/10.18421/TEM104-12

Houbrechts, M. (2024). *Using AI for Data Analysis: The Ultimate Guide (2024*). Available at: https://www.luzmo.com/blog/ai-data-analysis#:~:text=AI%20algorithms%20will%20search%20for,years%20to%20analyze%20big%20data

Ince, H., & Aktan, B. (2009). A comparison of data mining techniques for credit scoring in banking: A managerial perspective. *Journal of Business Economics and Management*, *10*(3), 233-240. https://doi.org/10.1016/j.jbankfin.2010.06.001

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the Lending Club consumer platform. *Financial Management,* 48 (4), 1009–1029. https://doi.org/10.1111/fima.12295

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence? *Discover Artificial Intelligence*, 2(1), 4. https://doi.org/10.1007/s44163-022-00022-8

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Kok, J. N., Boers, E. J., Kosters, W. A., Van der Putten, P., & Poel, M. (2009). Artificial intelligence: definition, trends, techniques, and cases. *Artificial intelligence*, *1*(270-299), 51.

Marinakos, G., & Daskalaki, S. (2017). Imbalanced customer classification for bank direct marketing. *Journal of Marketing Analytics,* 5(1), 14–30. https://link.springer.com/article/10.1057/s41270-017-0013-7

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *the AI Magazine*, 27(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine*, 27(4), 87. https://doi.org/10.1609/aimag.v27i4.1911

Olson, D.L., D. Delen, & Meng, Y.. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems,* 52(2), 464–473. https://doi.org/10.1016/j.dss.2011.10.007

OECD (2024), *Using AI in the workplace: Opportunities, risks and policy responses*, OECD Artificial Intelligence Papers, No. 11, OECD Publishing, Paris, https://doi.org/10.1787/73d417f9-en.

Paramesha, M., Rane, N. L., & Rane, J. (2024). Artificial Intelligence, Machine Learning, Deep Learning, and Blockchain in Financial and Banking Services: A Comprehensive Review. *Partners Universal Multidisciplinary Research Journal*, *1*(2), 51-67. https://dpi.org/0.5281/zenodo.12826933

Payne, E.M., J.W. Peltier, & Barger, V.A. (2018). Mobile banking and AI-enabled mobile banking: The differential effects of technological and non-technological factors on digital natives' perceptions and behavior. *Journal of Research in Interactive Marketing,* 12(3), 328–346. http://dx.doi.org/10.1108/JRIM-07-2018-0087

Post, H. (2023). *How Automation Helps Reduce Human Error and Improves Data Quality*. TRADESAFE. Available at: https://trdsf.com/blogs/news/how-automation-helps-reduce-human-error-and-improves-data-quality#:~:text=Automation%20can%20not%20only%20reduce,ultimately%20save%20time%20and%20resources

Redzik, J. (2024). *Największe banki w Polsce – styczeń 2024*. Bankier.pl, 2024-01-06 10:37. Available at: https://www.bankier.pl/smart/najwieksze-banki-w-polsce-styczen-2024

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2019). *AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence*, EUR 30117 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-17045-7, doi:10.2760/382730, JRC118163.

Situmorang, R. T. (2024). Ethics Related to the Adoption of Generative AI in Journalism. https://doi.org/10.30996/representamen.v10i02.10464

Smeureanu, I., G. Ruxanda, & Badea, L.M. (2013). Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management,* 14 (5), 923–939. http://dx.doi.org/10.3846/16111699.2012.749807

Soltani, M., M. Samorani, & Kolfal, B. (2019). Appointment scheduling with multiple providers and stochastic service times. *European Journal of Operational Research* 277 (2), 667–683.

Trivedi, J. (2019). Examining the customer experience of using banking Chatbots and its impact on brand love: The moderating role of perceived risk. *Journal of Internet Commerce,* 18 (1), 91–111. https://doi.org/10.1080/15332861.2019.1567188

ZBP & CPBiP (2020). *Sztuczna inteligencja w bankowości*. Wydawnictwo Centrum Prawa Bankowego i Informacji, Warszawa 2020

Artificial Intelligence (AI) in the Dentist's Work: The Perspective of Anthropomorphization and Cognitive Aspects of the Decision-Making Process

Guilherme Knupp Muniz, Luciana Paula Reis, Sérgio Evangelista Silva and June Marques Fernandes

DEENP, Federal University of Ouro Preto - ICEA, João Monlevade, Brazil

guilherme.muniz@aluno.ufop.edu.br
lucianapaula@ufop.edu.br
sergio.silva@ufop.edu.br
june@ufop.edu.br

Abstract: This research aims to evaluate the performance of an artificial intelligence (AI) tool called CranioCatch, which assists dentists by examining the anthropomorphization and cognitive aspects of the decision-making process. It remains unclear to what extent technology replaces or complements the work of healthcare professionals. Additionally, the specific cognitive aspects of the decision-making process attributed to CranioCatch, characterizing the anthropomorphization of the technology, have not been mapped. The methodology employed included "instruction to the double," "shadowing," and interviews. The study reveals how CranioCatch replicates cognitive activities performed by dentists by comparing, correlating, and associating images from a specific patient with similar images in a database. Through these steps (comparison, correlation, and association), which constitute the identification of the problem, it is possible to develop alternatives and select the best solution for the patient's situation. Thus, AI contributes to both the identification and action aspects of the decision-making process. From a physical perspective of anthropomorphization, the primary characteristic replicated by AI was the sense of vision. Regarding contributions to dentistry, the AI technology studied enhances the efficiency of professionals, notably reducing evaluation and diagnosis times. Due to its rapid and accurate analysis, AI is believed to significantly improve the work of healthcare professionals.

Keywords: Artificial intelligence, Radiographs, Decision-Making, Cognitive aspects

1. Introduction

The use of artificial intelligence (AI) technologies in healthcare is revolutionizing the work of professionals (Cordeiro *et al*., 2023; Reis *et al*., 2023). Due to demand, AI applications have been a rapidly developing field of study in dental radiology in recent years (Başaran et al., 2022; Reis *et al.*, 2023a). Considering the health sector, with a focus on dentistry, there are several AI technologies working, such as CranioWatch, an initiative that emerged from academic studies with the aim of evaluating dental x-rays (CranioCatch, 2023).

It is not known to what extent technology replaces or complements the work of healthcare professionals (Reis *et al.*, 2023b). Furthermore, it has not been mapped which cognitive aspects of the decision-making process are attributed to the CranioCatch technology, characterizing the anthropomorphization of the technology. Articles like Gorenstein (2023), Putra (2022), among others, reinforce the need for more studies and tests aimed at the use of AI in the healthcare area. Van Leeuwen et al. (2021) also mention that history has shown that validating these innovations in the real world is essential for decision-making. In this sense, this research seeks to evaluate the performance of an AI technology called CranioCatch, that operates in the dentist's work, from the perspective of the anthropomorphization and cognitive aspects of the decision-making process. Anthropomorphization consists of the incorporation of human abilities, including physical, cognitive and mental ones, into technical devices (Blut et al., 2021; Pelau, Dabija, Ene, 2021).

As there is a greater domain of cognitive characteristics involved in the decision-making process, a comparison of the activities performed by the AI with the dentist's work was carried out. The results aim to contribute to the use of AI, mainly in the dental area, from the moment in which the processes have been mapped and which characteristics are performed by the technology in facilitating and supporting the dentist's work.

2. Theoretical Reference

One of the most beneficial characteristics of AI is its ability to process significant volumes of data to provide results without the need for human action (Reis *et al.*, 2023a). Action which has proven to benefit multiple healthcare sectors that include bioinformatics, genomics, and image analysis (Gopinath, 2022). With the AI studies that have developed in recent years, a new opportunity for a solution has emerged in the diagnostic and image quality processes (Dicle et al., 2023). But, according to the literature, despite the performance of AI

with software such as CranioCatch (Başaran et al., 2022), U-Net network (Lin et al., 2021), among other algorithms based on deep learning (Kuwayama et al., 2019), (Shan et al., 2021), in which computers automatically learn from large sets of data in the field (Başaran et al., 2022), little is known about how it is changing the quality, efficiency, and costs of healthcare (Van Leeuwen et al., 2021).

2.1 AI in Dental Imaging Exams: CranioCatch Technology

AI is one of the most important technological developments that has caused deep changes in the life and production model of contemporary societies (Başaran et al., 2022; Cordeiro *et al.*, 2023; Cordeiro *et al.*, 2024). In recent years, AI research has rapidly emerged and developed in the field of dental and maxillofacial radiology (Putra, et al., 2022) and this can impact and add value to the healthcare community (KIM D., et al., 2022). Furthermore, AI studies in dentistry report that Deep Learning is promising in detecting many pathologies (Başaran et al., 2022). Thanks to the improvement of AI models in dental radiology, it is possible to predict the application of this technology by doctors in panoramic diagnosis and treatment planning (Başaran et al., 2022).

In dentistry, AI technologies have the potential to improve agility and effectiveness in diagnosis and clinical practice. To date, much research has studied AI applications in dental radiology, determining tooth numbering and detection, periapical pathology detection, caries detection, and restoration detection (Başaran et al., 2022). Table 1 presents some tools already on the market, according to the bibliography, and what their relationships are with the dentist's work.

Table 1: AI algorithms present in dentistry

Exam	Technology	Functionality	Benefits	Author	Aid x Replacement
Tomography	CNN model (deep learning)	Automated detection of chorioretinal diseases	Automated detection of macular diseases from Optical Coherence Tomography (OCT) images may be feasible using the CNN model	Kuwayama et al. (2019)	Improves work (replaces detection)
	U-Net network	Comparison of stock images	Perform automatic and accurate segmentation of the pulp cavity and tooth in cone beam computed tomography images.	Lin et al. (2021)	Assists with clinical work and tasks
Radiography	CranioCatch	Efficiently analyze 2D x-rays and quickly provide results to doctors	Detection of dental conditions on panoramic radiographs, except for cavities and dental calculus	Başaran et al. (2022)	It helps doctors in panoramic diagnosis and at the treatment planning
	THERE algorithms	AI in dental x-rays has shown great potential for a wide variety of purposes and can play an important role in assisting doctors in making decisions in the process	Current development of AI in dental imaging is not mature enough and requires substantial resources	Putra et al. (2022)	Assists doctors in the decision-making process.
Clinic	AI algorithms	Algorithms take unlabeled data and involve recognizing hidden data patterns that researchers may not have conceived of, producing unknown results. Principal component analysis and kmeans clustering are common methods used in learning.	Advancement in digitized technology for clinical examinations has made collecting healthcare data from dental patients less complex and cumbersome	Shan et al. (2021)	Assists with clinical work and tasks

Source: Prepared by the authors.

The technologies present on the market vary in the scope of their operation. Table 1 was organized so that the lines represented the different technologies. It is possible to observe the specific areas in which each technology operates, the main benefits presented in the literature and their relationship with the

professional's work, that is, whether the technology is assisting in the work or is replacing some function of the worker.

In the present work, the technology studied was CranioCatch, due to its per-positive and successful training in x-ray analysis. This technology has great potential to contribute to the dentist's work. The AI model is based at the Convolutional Neural Network (CNN) deep learning method. It proposed to detect ten different dental situations, rated as successful in detecting dental conditions on all panoramic radiographs except for caries conditions and dental calculus (Başaran et al., 2022). In dentistry with CranioCatch, solutions will improve the quality of healthcare, allowing patients to be treated faster and more accurately. This software selects the images to be analyzed and loading them into the system, returning the x-ray with the diagnosis in seconds. In this sense, dentists will be able to provide a better service to patients, obtaining results more reliable in x-ray analysis (CranioCatch, 2023).

2.2 The Relationship Between AI and Cognitive Aspects of the Decision-Making Process

When it comes to decision-making, AI emulates the cognitive abilities of human beings, in the sense of automating some decisions (Gopinath, 2022). Research into AI began shortly after the emergence of modern computers, with the goal of building intelligent "thinking" machines. Since the birth of AI, interactions between it and brain science have been carried out (Fan et al., 2020). Still according to Gopinath (2022), there are two ways in which neurosciences inspire AI, one being its ability to imitate human intelligence and the other ability to develop a neural network, which can imitate the structure of the human brain.

A decision is defined as a moment in a continuous process of evaluating alternatives related to an objective, in which the decision maker's expectation regarding a certain course of action impels him to select (Harrison, 1981). The beneficial relationship between AI and neuroscience is reciprocal, and AI has currently been widely investigated in the context of neuroscience. In this sense, AI in different situations produces results equivalent to those of human beings, emulating human intelligence. Therefore, a simple way to develop AI is to combine it with brain science and related fields such as cognition science and psychology (Fan et al., 2020). The literature presents some theories that describe the decision-making processes, and the cognitive aspects present in decision-making. Table 2 presents these theories.

Table 2: Elements of decision-making

THEORY	COGNITIVE ASPECTS	DESCRIPTION	AUTHORS
Decision-making process	Assess and determine	Ascertain the truth, determine a course of action, and persuade.	Barnard (1938)
Decision-making process on bounded rationality	Intelligence, design and the choice	Intelligence: Search, interpretation and supply of factors relevant to the decision, such as data and information. Design: Choice of methodology to be used or alternative paths that bring potential gains. Choice: Option for a line of action involving judgment and values, implying the renunciation of other alternatives.	Simon (1965)
Decision-making process	Recognize, identify, evaluate and choose	Recognition of a situation that requires a decision; Identification and development of alternative courses of action; Assessment of alternatives; choosing one of the alternatives and implementing the selected course of action.	Newman (1971)
Strategic decision process	Identification, development, selection	1- Opportunities, problems and crises are recognized and relevant information is collected and problems are identified more clearly; 2- Alternative solutions to problems are generated and modified; 3- The alternatives are analyzed and scanned and an alternative is chosen.	Mintzberg (1976)

Source: Prepared by the authors.

From the elaboration of Table 2, achieving a visual model that allows comparison between different theories, it is possible to identify which are the most common cognitive aspects present in the decision-making process. Authors such as Simon (1965) and Mintzberg (1976) theorized decision-making by dividing it into phases, basically the first two being associated with information collection and identification and the last phase linked to the action to be performed and information processing. Based on the relationship between such theories, it is possible to divide the decision-making process into two main stages: information collection and information processing.

3. Methodology

This study was divided into two stages. After analyzing the literature and websites that address the technology of CranioCatch, the first stage began with data collection. The dentist´s work process was mapped using data collection and checking methods: 1 - “instruction to the double” and 2 - “shadowing” After data collection, the second stage began. At this moment, the work process flow was designed and activities that could replace or complement the work of the health professional were identified.

The method of “instructing the double” emerged in the 1970s, during the Italian Workers Movement (IWM) with the doctor Ivar Oddone (Pereira, 2017). In order to provide a better understanding of the processes between interviewee and interviewer, the proper technique involves promoting the interviewee in a position of educator. The interview was conducted remotely, via Google Meet for approximately 50 minutes. A sequence of questions was asked about how treatments were carried out, how anomalies were identified in examinations, whether it was possible to observe some type of pattern and how artificial intelligence could act in the area.

The "instructing the double model" should avoid questions that express motives or reasons such as “why?”, so that the interviewee does not feel coerced, which would lead to the omission of some information for fear of incorrectly speaking about a process. Questions like “how to do it?” and “what should never be done?” are prioritized. All processes were noted while the professional responded, with more details being requested in some situations. After completing the interview, a work process flow was created that would be duly validated after the shadowing analysis.

With the intention of corroborating the responses obtained, the “shadowing” technique was used, a research technique that involves a researcher closely following a member of an organization over an extended period (McDonald, 2005). By having the action processes mapped, and with the authorization of the professional and the patient, it was possible to monitor some services to analyze in practice what was being done and how the professional's decision-making was made. Two patients were monitored, whose consultations lasted around 30 minutes, and who had already undergone the x-ray procedure. It was possible to monitor how the professional identified the injuries (using the senses, such as touch and vision, and clinical experience), how the report was provided and how the treatment was conducted. All analyzes carried out in this data collection stage were being meticulously compared with what had been reported by the dentist during the first interview, thus being able to validate the previously prepared work map.

The focus of the questions and follow-up was to understand as much of the work processes as possible, to obtain a stage map. Therefore, the creation of a visual analysis model would help in understanding how each process is carried out, and what is necessary, such as: experience, use of the senses and comparisons, to obtain a diagnosis and propose a treatment. Upon finalizing the workflow map, it was presented to the professional for final verification. The analysis consisted of comparing the data obtained in both methods with the performance of CranioCatch presented in the bibliography and pointing out in which fields of work the technology is applicable.

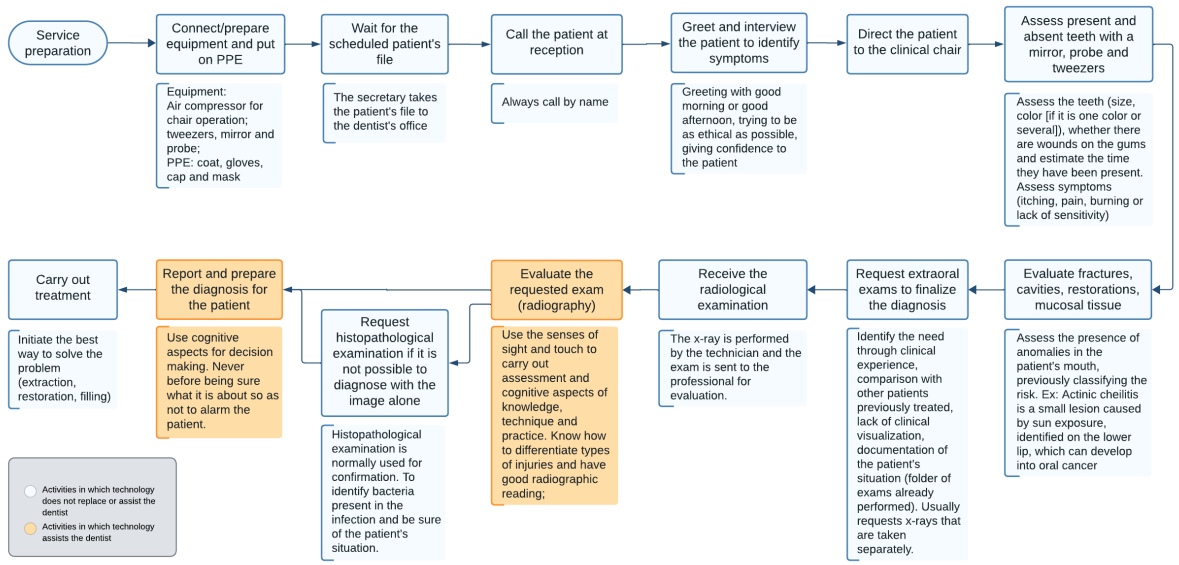
4. Data Analysis

In dentistry, the professional's work includes physical and mental activities that, combined, can generate a complex working day.

4.1 Dentist´s Workflow: Identification of Activities That can be Replaced or Complemented by AI

To identify where AI is used to replace or complement human work, a mapping of the activities involved in the dental patient care process, as shown in Figure 1, was created. All mapped actions come from the professional's perspective. After data collection was completed, the workflow model was developed. All the

processes described were highlighted by the interviewed professional, from calling the patient to the room to presenting the report. Each box represents an action performed during the activity.



Source: Prepared by the authors

Figure 1: Dentist´s workflow

The workflow analysis consisted of identifying which processes/actions could or are already replaced by AI technologies. The light blue boxes represent activities that, in principle, could not be replaced by AI. Just the radiography and emission assessment activities of the report were considered as possible to have the assistance or support of AI, highlighted in orange in the flow image. In a dentist's work, with an emphasis on image analysis, the evaluation of x-rays can take up a lot of the professional's time, and it is possible for some details to go unnoticed.

4.2 Anthropomorphization of CranioCatch Technology

The decision-making process was evaluated in two parts: the first which addresses the collection of information from the senses and the second which corresponds to the processing of this information collected from cognitive aspects. Regarding the first, it was observed that of the five senses, the CranioCatch technology only explores “vision”, in the sense of perceiving and visualizing the image of an x-ray. So, after “viewing” this image, it starts processing this information. Thus, the cognitive aspects reproduced by technology, in the analysis of x-rays to facilitate decision-making by the professional, are represented in Table 3.

Table 3: Anthropomorphic aspects of the CranioCatch

CRANIOCATCH DECISION-MAKING PROCESS ASPECTS				
IDENTIFICATION			**ACTION**	
COMPARISON	**CORRELATION**	**ASSOCIATION**	**DEVELOPMENT OF ALTERNATIVES**	**SELECTION**
Compare the image with other images to analyze similar situations	Identify similar images (relate images) to predict/outline diagnoses	Associate nearby images with atypical situations, from which there is knowledge about what type of anomaly it may be	Once the diagnosis has been defined, identify alternatives to solve the problem: complementary actions are defined to complete the diagnosis and/or immediate intervention.	Selection of the alternative that best suits the patient's situation.

Source: Prepared by the authors.

The preparation of the table consists of presenting which activities the CranioCatch software performs, in order to reproduce cognitive activities carried out by dentists. In this case, it executes comparison, correlation and association with nearby images to identify diagnostic possibilities. Finally, he prepares a report. The

evaluation action and result of the technology takes place in seconds (CranioCatch, 2023) enabling greater agility in work and providing support in professional decision-making.

5. Discussion

Considering the limited understanding of the extent to which technology replaces or complements human work, particularly in the context of healthcare professionals. Furthermore, there were no maps of which human aspects are attributed to the CranioCatch technology, characterizing the anthropomorphization of the technology.

It is an AI trained to evaluate imaging exams, in which categories of situations are identified. The stages of developing alternatives and selecting the result of the CranioCatch technology involve image selection, analysis (comparing, relating and associating the new x-ray with images from your database) and presentation of the result.

With its focus on the area of dentistry, the AI technology studied presented characteristics that help the professional's work, reducing evaluation and diagnosis time. Technology assists in carrying out diagnosis and evaluating treatments that may occasion better treatment results. Other positive aspects are the greater safety for the patient, increased production and reduced service time due to its capacity to analyze large volume of data related to different patients comparing situations at different stages of the treatment process.

CranioCatch reproduces anthropogenic activities generating diagnosis and is classified as successful in detecting dental conditions on all panoramic radiographs. Once elements of the theory of decision-making processes were applied to AI, the reproduction of cognitive aspects was observed, mainly in the identification phase. Based on theories that explain how it is carried out and what the stages of decision-making are, mainly on Mintzberg's theory (1976), it was possible to identify and map which cognitive aspects the technology reproduces. A further step towards understanding the extent to which technology assists or can replace human work, so far, in radiographic analysis. However, the possibility of being used in clinical analysis and diagnoses in the future cannot be ruled out.

Other studies have sought to evaluate the use of AI in healthcare work processes (Reis *et al.*, 2023a; Van Leeuwen et al., 2021; Shan et al., 2021). However, this study at the level of AI applied to the work of dentists which use the sense of vision for recognition and collection of information to cognitive aspects for processing the information collected. The main contribution of the research is to analyze the relationship of AI in decision-making in a dentistry work process, surveying the cognitive aspects of decision-making. From this, assessing how an AI partly replaces and complements this work.

6. Conclusion

Technology assessment CranioCatch from the perspective of anthropomorphization in the dentist's work, made it possible to verify situations in which the use of AI can replace or complement the dentist's professional activity, in the evaluation of x-ray exams.

The technology acts cooperatively in the professional's work, supporting decision-making with a quick analysis of the x-ray. Relating the processes carried out by AI with decision-making theories contributes to studies related to AI, especially in the health area.

As limitations of this study, it is impossible to verify performance of the software in the evaluation of radiography and the comparison of efficiency between the dentist's evaluation. Since there were no statistical tests for comparison. Furthermore, there is a lack of the patient's existing clinical knowledge, which may impact the outcome of the report. Once during the interview carried out with the dentist to obtain data, it was emphasized by her that the patient's clinical record and the examinations and treatments carried out in the past are linked to the evaluation and the result.

For future work, it is suggested to conduct interviews with more professionals, including those who use technology to obtain feedback on how useful it is in their daily lives. It is also suggested to carry out statistical tests to compare the efficiency of AI and dentist assessment.

Acknowledgements

The authors would like to thank the Universidade Federal de Ouro Preto (UFOP/Brazil) (www.ufop.br), Programa de Pós-Graduação em Engenharia de Produção (PPGEP-UFOP), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, grand number APQ-00951-22), Coordenação de Aperfeiçoamento de

Pessoal de Nível Superior (CAPES) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), for the support and funding during the development of the research.

References

Barnard, C.I. (1938). The Functions of the Executive, Harvard University Press, Cambridge, MA.

Başaran, M., and Other. (2022). Diagnostic charting of panoramic radiography using deep-learning artificial intelligence system. *Oral radiology*, 1-7.

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, 49, 632-658.

Cordeiro, R. F., Reis, L. P., & Fernandes, J. M. (2023). A hierarchical model for industry 4.0 concepts. RAM. Revista de Administração Mackenzie, 24, eRAMR230061.

Cordeiro, R. F., Reis, L. P., & Fernandes, J. M. (2024). A study on the barriers that impact the adoption of Industry 4.0 in the context of Brazilian companies. *The TQM Journal*, 36(1), 361-384.

Craniocatch. Artificial Intelligence Solutions in Dentistry. 2023. Available at: <https://www.craniocatch.com/en/>. Accessed on: 11/30/2023

Dicle, O. (2023). Artificial intelligence in diagnostic ultrasonography. *Diagn Interv. Radiol*, 29, 40-45.

Fan, J., and Other. (2020). From brain science to artificial intelligence. *Engineering*, 6(3), 248-252.

Gopinath, N. (2023). Artificial intelligence and neuroscience: An update on fascinating relationships. *Process Biochemistry*, 125, 113-120.

GOPINATH, Nishanth. (2022). Artificial intelligence and neuroscience: An update on fascinating relationships. Process Biochemistry.

Gorenstein, L., and Other. (2023). AI in radiology: is it the time for randomized controlled trials?. *European Radiology*, 33(6), 4223-4225.

Kim, D., and Other. (2022). Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. *Nature communications*, 13(1), 1867.

Kuwayama, S., and Other. (2019). Automated detection of macular diseases by optical coherence tomography and artificial intelligence machine learning of optical coherence tomography images. *Journal of ophthalmology*, 2019.

Lin, X., and Other. (2021). Micro–computed tomography–guided artificial intelligence for pulp cavity and tooth segmentation on cone-beam computed tomography. *Journal of Endodontics*, 47(12), 1933-1941.

McDonald, S. (2005). Studying actions in context: a qualitative shadowing method for organizational research. *Qualitative research*, 5(4), 455-473.

Mintzberg, H., Raisinghani, D., & Theoret, A. (1976). The structure of" unstructured" decision processes. *Administrative science quarterly*, 246-275.

Newman, J. W. (1971) Management Applications of Decision Theory.

Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122, 106855.

Pereira, M. D. S. (2017). Movimento operário italiano, Ivar Oddone e a instrução ao sósia. *Trabalho & Educação*, 26(3), 13-27.

Putra, R. H., and Other. (2022). Current applications and development of artificial intelligence for digital dental radiography. *Dentomaxillofacial Radiology*, 51(1), 20210197.

Reis, L. P., and Other. (2023). Innovation in Ophthalmology and Dentistry Services: Benefits and Challenges of using AI. *In ECIE 2023 18th European Conference on Innovation and Entrepreneurship Vol 2.* Academic Conferences and publishing limited.

Reis, L.P., Fernandes, J.M., Silva, S.E. and Andreosi, C.A.d.C. (2023). Managing inpatient bed setup: an action-research approach using lean technical practices and lean social practices. *Journal of Health Organization and Management*, Vol. 37 No. 2, pp. 213-235.

Shan, T., Tay, F. R., & Gu, L. (2021). Application of artificial intelligence in dentistry. *Journal of dental research,* 100(3), 232-244.

Simon, H. A. (1965). Administrative decision-making. Public Administration Review, 31-37.

Van Leeuwen, K. G., and Other. (2021). How does artificial intelligence in radiology improve efficiency and health outcomes?. Pediatric Radiology, 1-7.

AI in the Learning Environment: Examination of Pedagogical, Psychological, and Ethical Implications

Gala Krsmanovic[1] and Fadi Deek[2]
[1]Department of Humanities and Social Science, New Jersey Institute of Technology
[2]Department of Informatics, New Jersey Institute of Technology

gk72@njit.edu
fadi.deek@njit.edu

Abstract: This paper provides an overview of how AI tools and capabilities are being integrated in teaching and learning. It also addresses ethical aspects of integrating AI and other advanced technologies in these processes as well as the resulting implications. These issues include algorithmic fairness, privacy and surveillance, intellectual property rights, misinformation and violence, health concerns, and social implications. In addition, how access to such tools will impact the cognitive functions of their users, particularly learners, has become a relevant question. Much more information is now becoming easily accessible to a wider population, making it important to understand and mitigate AI ramifications to human learning, reasoning, and perception, especially for young learners whose brains are still in the early developmental stages. To that end, this paper also reviews the current literature regarding the impact of AI integration in education, focusing on the potential effects this can have on the cognitive abilities and functions of learners. Thus, with AI making its way into the educational realm, from pre-college to advanced studies levels, understanding the resulting pedagogical, psychological, and ethical ramifications can guide thoughtful development and appropriate integration of AI tools in support of the learning environment.

Keywords: AI ethics, Societal impact, Learning, Cognition, Regulation

1. Artificial Intelligence in Education

The speed of technological innovation and acceptance is accelerating. Resulting technological systems are embraced and integrated into the educational processes, with sophisticated AI tools capable of general-purpose language understanding and content generation becoming available to learners of all ages. It is certain that the educational system will fundamentally change as a result, making it necessary to examine, understand, and mitigate the challenges we are certain to face, including how such tools will impact the cognitive functions of their users, specifically learners. It is a fact that AI applications, especially language models, have been adopted quickly and broadly in educational settings, and are commonly used by instructors and learners for their coursework. These tools help educators develop and deliver course content, assess students' knowledge and progress toward their learning goals, and provide targeted feedback and timely support to improve learning outcomes.

While some types of AI systems had already been in use in academic settings, large language models capable of understanding and generating text and other content are allowing to drastically reimagine education. This includes access to things like algorithmic learning systems, predictive analytic tools, automatic assessment, virtual reality, and more. Specifically, promising personalized learning tools, with machine learning algorithms that analyze learners' performance and create customized lessons, aligned with individuals' strengths and weaknesses, are receiving wide acceptance. Virtual Reality devices allowing for better understanding of abstract concepts or practicing complex procedures are increasingly in use. Pedagogies aiming to make learning more efficient while rethinking the important role of the human component in the process are also underway.

The advantages for learners, teachers, and organizations are as indisputable as the drawbacks. In addition, while many have already forecasted that this can also increase motivation and engagement, others warn of risks and consequences, such as diminished capacity for critical thinking, declined memory, reduced attention span, and lack of motivation. Furthermore, there are concerns that it is becoming difficult to ascertain authenticity of intellectual creations, and therefore academic integrity. While some have gone as far as to argue in favor of severely curtailing or forbidding the use of such tools in schools, others see this as an opportunity to guide learners toward ethical and responsible conduct in all matters of academic work. This makes it important to consider the potential ethical and social challenges of these developments as well as the responsibilities that institutions, educators, and students face.

1.1 Generative AI

Probably the most widespread AI tool that has taken over the education system is Generative AI. They are making it possible to create original content in a matter of seconds, for which we would need hours or days to

make, by replicating our own creative processes. The possibilities for this are endless. However, it is important that we ask the question whether we should accept AI as the new "normal" and embrace it as a benefit, or we should be concerned about where it will lead us in the future.

Generative AI models have a long history, going back to the 1960s. They are now used in chatbots to generate human-like responses. Recent advances in machine learning and natural language processing, and the release of new Generative AI systems, have made it possible to bring them into everyday applications. The technology was brought to the public when ChatGPT came out on November 30, 2022. It quickly became a favored platform for creating content for everyone, from students in need of a school essay to employees using it to complete their work.

Generative AI will continue to become increasingly popular and, consequently, the more it develops, the more questions will arise. It is reasonable to ask whether homework assignments will even make sense anymore now that students have discovered AI models; or, how might employers differentiate workers and reward them according to their work contributions given the use of such programs.

Generative AI enables individuals and organizations to do things that were not previously possible, both in terms of speed and scale. How it will eventually be exploited is a fundamental point. The public introduction of Generative AI is, in many ways, a first ever test for us, as a human race, as to how we will react to such a wide- and far-reaching innovation. For example, how might students change when they are given an easier path to meeting their responsibilities, whether it is to satisfy a curiosity or something more important; and how some students may completely give in to technology and allow it to do their work, while others use it as an aid to do a better job and be more productive. In fact, the whole notion of work and productivity is being re-written. Take software engineering, as an example, where substantial and sophisticated code can be quickly and readily generated, based on natural language directives, facilitated by the enormous volume of open-source code used by machine language models for their training.

Furthermore, since Generative AI systems make new patterns from countless examples that already exist on the web, the question is who should own the generated content. Since Generative AI products are not based on one source, but on an aggregate number of them, this type of infringement might not be easily noticeable, but it is certainly real and could be discerned. On one side of the debate, there is a notion that intellectual rights belong to the AI model that created the material. Or possibly to an original creator whose work is on the internet, even though it is essentially impossible to make specific attributions. This is because the algorithms are taught to recognize patterns and relationships in large data sets and generate new content. Since the generated material is created using such data, the creators of this work are asking for its copyright. Official US and European laws currently state that AI cannot own the copyright as it is not an author and does not have a legal personality.

1.2 Deepfake Videos

Deepfake Video Technology—videos created using deep learning AI methodology and technology allowing the likeness of individuals in a video to be swapped with any other person's likeness. Deepfakes have already been integrated in education, as a means of visualizing work. While this technology can make some aspects of education more appealing especially to younger learners, perhaps by making a video of a historical figure speaking about an event in the past rather than having a teacher say it, it does also raise some serious concerns.

The reason deepfakes are so convincing, more than a photoshopped picture or an altered audio recording, is that the human brain responds the most to visual communication stimuli compared to other sensory information, a psychological phenomenon known as the Colavita visual dominance effect (Koppen & Spence, 2007), particularly when it involves non-verbal communication such as body movement. Because of this, it is especially hard for us to recognize a real video from a fake. The more we are aware of the fact that videos can be photoshopped, just like images can, the more deliberate we can be in choosing what to believe. This is of utmost importance especially with young learners. In addition, while deepfakes are legal, it is important to understand that some content could be considered to have breached legal codes and is thus liable. Clearly though, whether we talk about public concerns, such as law and order, or personal ones, like reputation and social identity, we are starting to approach an era where protection from technological attacks is becoming a human right.

The best approach to countering any type of deception is through education. The more we know, the less impressionable we become. However, that can bring on another problem, the interference with our ability to acquire knowledge. Knowing that anything can be altered, our brains can doubt every piece of media we see, even the ones that are legitimate, and can lead us to acquire false beliefs. This is known as the epistemic threat

of deepfakes. When this becomes the case, videos will no longer be reliable, especially as trusted evidence, or fair sources of information. This is what makes deepfakes controversial, and diminish its promise for teaching and learning. However, one thing is certain, there is a need for more research into this technology and laws to regulate it.

1.3 Video Gaming

Video games have gained immense popularity in the last decade. As such, significant efforts and resources are being invested to improve the gaming experience. In addition, video games are more accessible than ever. There is no need to go to the arcade room or buy a gaming console anymore. Video games have evolved significantly from their early days. It is fair to say that gaming is here to stay as evidenced by the unprecedented surge in popularity and profitability the industry is enjoying. Also, culturally, gaming is becoming more mainstream. Not only is there a noticeable and rapid increase in the number of people playing games, they are doing it both earlier and later in life.

The introduction of AI into gaming has made it more creative and realistic. Thus, it is also easier than ever for any individual, of any age, to enter the world of gaming and perhaps be exposed to the dangerous side of this world. In fact, AI is enabling the creation of a fully digital universe, the Metaverse, the 3D virtual world, where people interact, socialize, and engage in various activities through virtual environments. These virtual worlds are reshaping societies, economies, and even human identity. That being said, an educational metaverse, where students and teachers interact and learn, is something we will see in the future.

Metaverse offers a glimpse into the possibilities in the converged physical and virtual spaces where digital facsimiles of ourselves, avatars, move freely from one experience to another, taking our identities and assets with us. It is expected that community participation and collaboration will produce innovations and user-generated content for this space. It is not unreasonable to forecast that these exciting developments will only amplify the challenges, particularly ethical ones. While these advancements have made the experience of gaming more realistic and exciting, they are also responsible for some complex dilemmas. It is relevant to mention that the World Health Organization has a classification for a “Gaming disorder”, defined, in part, “as a pattern of gaming behavior characterized by impaired control over gaming, increasing priority given to gaming over other activities to the extent that gaming takes precedence over other interests and daily activities, and continuation or escalation of gaming despite the occurrence of negative consequences.”(Darvesh et al., 2020) However, gaming is also credited for promoting better cognitive performance among its users (Bork, 2012).

2. Emerging Implications and Ethical Ramifications

A major issue with AI is the transparency and accountability of its decision making. Because of the complexity of their design, making AI systems difficult to understand. In addition, there are a host of other matters surrounding the development and deployment of AI tools that require attention. Among the most serious are fairness of predictive algorithms, that is the principle of designing AI algorithms in a way that does not discriminate against individuals or groups. Also, privacy, security, and safety are valid issues given the personal information being collected and stored by corporations, organizations, educational institutions, and governments and how it is used, including the sharing or selling of data as well as the risk for monitoring of individuals that lead to loss of privacy. Furthermore, AI-generated work and impact on intellectual property rights is an issue that is still to be resolved. Misinformation, the potential for violence, addiction, and health problems are also significant concerns. These issues are even more pronounced in education given the different age groups and diverse population potentially impacted.

2.1 Predictive Analytics

Personalized learning is an advantage intended to be a means that enables instructors to tailor unique learning goals based on individual needs. It relies on adaptive algorithms to create content and experiences for learners appropriate to their own profiles and accomplishments based on collected user data. Algorithmic predictions in machine learning use sets of micro- and macro-data to form models that can subsequently be applied to people and objects, and predict outcomes based on previous experience. This type of automated, or semi-automated, decision-making is favored because predictive algorithms are capable of processing much more information, and much faster, than we can and thus impacting matters of significance in domains such as education, but also with potentially wider implications in other domains. For example, the potential of infusing bias into student assessment approaches because human decision-making is being replaced by algorithms of intelligent machines that are subject to biases inherent in the data used to train these algorithms.

The ability to derive insights from multiple data sets and make impactful decisions through algorithms may result in unintended effects on certain individuals and groups that have already shown to be potentially unfair with a risk for racial, gender, or other biases. The predictive algorithms are not thinking, they are making mathematical calculations based on the data that we, humans, provide them and others they collect.

2.2 Algorithmic Fairness

A requirement that AI algorithms are free from bias and do not reinforce existing social discriminations seems reasonable and logical but, at least for now, remains elusive. Decisions commonly and routinely made by humans, which benefit from the innate normative moral sense we possess, despite variation across cultures and communities, are progressively made by computer algorithms that can reach incorrect conclusions. As a result, new problems that have real-life implications are starting to emerge. AI algorithms learn from the data they are trained on, and if that data is prejudiced, the algorithms will unfairly affirm this bias. This requires careful consideration of the data used to train the algorithms, the foundations of the model, and the measures to assess its performance.

In the pre-AI era, humans made simple and complex decisions using available data combined with their instincts, beliefs, and values. Now, algorithms are increasingly taking on some of these decision-making processes because it is believed that they are faster and more accurate in many situations than humans. However, it is important to contemplate these serious matters, considering that AI, as much as it has evolved in the last decade and how much it will still evolve in the future, is still artificial. The fact remains that AI is unable to consider moral judgment in making its calls, and thus can reach incorrect conclusions and generate misguided decisions that create unfair outcomes, such as privileging, or disadvantaging, some people over others.

As technology advances, approaches for mitigating and eliminating these faults can be developed. Even so, the question remains whether a machine can ever be completely accurate, fair, and follow prevailing societal norms considering that AI is, in large measure, based on identifying patterns, and that implicit and explicit stereotypes are essentially a type of pattern by their nature. At the same time, efforts to identify, eliminate, and prevent bias, in all forms, caused by algorithms of personalized learning tools, that are being addressed through audits and checks have yielded acceptable results. Lasting and effective solutions to these problems are possible with social awareness and governmental regulations.

2.3 Privacy and Surveillance

As AI continues to pervade many aspects of our personal and professional lives, it becomes easier and quicker to gather, store, analyze, and share information in ways that intrude on privacy. This brings up a number of concerns with its use in education. Starting with privacy and surveillance, AI devices that track students' location, attention, and behavior capture and store sensitive data. Determining data ownership and control can be complex, especially when data is collected by third-parties. Observing students using technology for seemingly legitimate reasons, such as identifying those who may be inattentive or distracted during class, can limit creativity, or freedom, and amounts to surveillance. This authoritarian approach to data gathering also raises the question of profiling, violating privacy rights, and may lead to potential exploitation. There is also a rising trend in using AI facial recognition tools for predictive profiling by tracking and monitoring, sometimes excessively, the actions and movements of students resulting in incidents of bias and discrimination.

Nearly all processes students engage in on a routine basis requires the disclosure of some personal data such as name, address, phone number, date of birth, and gender. Complex processes require the collection of much more sensitive forms of personal data points that require a greater level of protection such as social security number, racial or ethnic background, and biometric or genetic information. Much more intrusive data requests can be made depending on the nature of the action triggering it. This is becoming increasingly pronounced in our schools.

The risk of individual harm, especially troubling for minors, are examples of the potential consequences stemming from unauthorized data breaches. Making this more unsettling is the fact that it is not clear who owns the data collected from surveillance systems, how securely it is stored, and with whom it is shared, let alone whether appropriate consent was obtained to collect the data in the first place. It can be further argued that it is difficult to safeguard impacted individuals, whose images and activities are being captured, often unaware, by cameras becoming omnipresent in classrooms, a technological intrusion that is not transparent.

There are other, more serious, data privacy questions regarding digital surveillance, which are necessitated by the need to protect schools and their occupants, including students, teachers, staff, and visitors. That is the use

of the technology to identify individuals without their informed consent or knowledge especially by identifying and analyzing biometric traits of the face, voice, gait, and other features of a given image. There are legitimate algorithmic concerns in such situations. Relying on machine learning in the public safety sphere amounts to predictive profiling, which essentially requires the AI to anticipate that crimes will occur, who might the offenders be, and who might the victims of these potential crimes be. Although not purposeful, given the questions of algorithmic fairness and bias in training data, implications to individuals and groups may disproportionately affect certain sectors of society more than others.

2.4 Intellectual Property Rights

There are many moral questions, beyond plagiarism and copyright, and it is crucial that we are able to define the limit to which an action is ethical or not, especially considering what the future of AI still holds. Ever since Generative AI gained popularity, the development of software detecting AI-generated material has begun. From a pragmatics perspective, it is less important to know whether or not the material can be detected as generated by AI. The question should be which of our actions are likely to be unethical and violate certain norms, not just in education, but also in everyday lives.

The issue of potential academic misconduct is the simplest to address. As it has always been with such matters, instructors' philosophy and preferences have created a range of reactions, from a very vigilant approach to stamping out the use of AI-generated content, on one hand, to a more tolerant approach of accepting such generated material, on the other. It seems reasonable, for a writing exercise, that generated content can be used as a starting point to inform one's work, brainstorm ideas, and prepare notes, just as can be done with textbooks, library resources, and web materials; of course, with appropriate citations and attributions. Obviously, there should be no tolerance for plagiarism regardless of how it occurs.

An ethical concern also arises whether authors of intellectual work used by Generative AI are being disadvantaged. Normally, if a person's work is used in any form, the owner, author or artist, would have to be referenced or otherwise compensated, thus receiving credit for their original creation. However, Generative AI does not do any of these things, raising ethical flags among writers, scientists, artists, and other content creators.

2.5 Misinformation, Violence, Health Concerns, and Social Implications

Digital harassment, hate speech, and fake news are becoming more prevalent across societal domains including education. AI is now an unwitting enabler that carries a big risk of spreading misinformation or using the generated material for misconduct. Such behavior is already a major concern in schools as it is easily effected through AI in the form of threatening messages and abusive language, disclosing unauthorized private information or images, and disseminating erroneous information and spreading rumors, creating false profiles or assuming social media identities of others, and making fake videos or AI-generated content that is difficult to distinguish from real ones. Deepfakes, for example, especially those of sexual nature, have been harmful to the reputation and well-being of students subjected to such behavior. The use of video gaming as a pedagogical tool has added to this complication and their integration in education requires careful planning. Virtual worlds like the Metaverse are anticipated to undergo further growth, offering even more immersive experiences. The Metaverse, in particular, has the capacity to profoundly impact the education domain and, as it continues to evolve, could reshape how we interact, learn, and engage with technology in ways we have yet to fully comprehend.

The American Psychological Association indicates that a high percent of video games currently on the market contain some form of violence, while acknowledging that there is not a direct connection between playing these games and being violent. However, a psycho-social theory known as the General Aggression Model suggests that exposure to violence through video games creates cognitive scripts of aggression in the human brain.

Additionally, AI-powered games collect and analyze vast amounts of player data to create personalized experiences and optimize in-game features. However, this comprehensive data collection presents an ethical challenge regarding data privacy–how is this information used, stored, and shared. Similarly, there are social concerns arising from the prevalence of gender stereotyping in games. For example, only a small percent of video games feature a female protagonist. When female characters are included, they often are sexualized, characterized by revealing clothing and exaggerated body proportions. Finally, there are concerns that some video games can be used maliciously, for example to normalize deception or promote suicidal ideation. Naturally, the aim should be to foster positive and inclusive communities, and resist the temptation to propagate biases, toxicity, harassment, stereotypes, or discriminatory behavior.

3. Impact of AI in Education on Cognition

Interest in how technology impacts human cognition is growing with the rise of AI tools capable of language understanding and generation. Reliance on AI tools, instead of thinking processes raises concerns on how AI will affect cognitive development, especially in children and youth. There already have been some psychological phenomena documented when it comes to the relationship between the brain and digitization. Specifically, the tendency of not committing valuable information to our memory because we know that it is easily accessible online, known as "Digital Amnesia" (Sparrow et al., 2011). Understanding how this technology affects our cognitive ability and mitigating ramifications to human learning and reasoning is crucial for pedagogical purposes.

Although theoretical models to explain potential impacts and consequences have been proposed, only a limited number of experimental studies have been conducted so far, due to the recency of AI (making it impossible to observe longitudinal effects or notice neurological changes) and methodological challenges regarding validity and reliability of measurement tools. Therefore, most current experimental literature in this area relies on questionnaires, self-reported measures, and pre- and post-test designs in groups with or without engagement with AI tools. For instance, investigating the influence of Generative AI on critical thinking skills in a sample of 107 graduate students (Essien et al. 2024) found that students who were allowed to use ChatGPT improved basic cognitive abilities compared to those who were not allowed, but there was no significant effect on higher-level skills like creativity. This suggests that while ChatGPT can assist with fundamental cognitive skills, its effect on higher-level functions remains limited.

Similarly, another study on 637 students found that AI improved students' efficacy and engagement, but had no significant effect on critical thinking skills (Jia & Tu, 2024). This increase in self-efficacy may be due to the psychological boost students receive from feeling more capable of handling academic challenges. Conversely, a comprehensive two-phase study from 2021 focused on the relationship between inputs, such as the time invested, and outputs, such as learning, knowledge transfer, and skill adoption, when students use or do not use digital technology in their education The results found that the use of digital technology in higher education has no positive impact on students' efficacy and the cognitive outputs, and that students are actually more efficient in their higher education goals when they do not use digital technology (Lacka et al., 2021).

The cognitive process behind overreliance on AI during decision-making has also been investigated (Bucinca et al., 2021). It was hypothesized that people rarely engage in critical and analytical thinking when deciding to trust the information given by AI. The experiment was conducted on 199 people divided into three groups, each engaging with one of the three cognitive forcing designs, and compared it to a control group. The results showed that cognitive forcing significantly decreases overreliance on AI. The importance of this study lies in the fact that overreliance is one of the main causes of critical thinking decline when engaging with AI, which poses a significant danger to students. It also represents a valuable way to determine if a student simply plagiarized the content generated by ChatGPT or only used it as a tool and analytically thought through the information that was generated.

3.1 Attention and Motivation

The most commonly investigated cognitive function impacted by technology is attention, likely due to the constant stream of stimuli we encounter (e.g., common notifications, working in multiple tabs, seconds-long videos on social media, etc.). As the intensity of these stimuli increases, the level on which we interact with them decreases, a behavioral pattern called "media multitasking".

While early researchers suggested that multitaskers would excel in attentional task-switching (because they "practice") results actually showed the opposite, and analysis suggested that those who engage in frequent and extensive media multitasking show worse attention abilities because they are too susceptible to irrelevant environmental stimuli and distractions (Ophir et al., 2009). These results were also supported by MRI imaging of the brain observed while subjects performed attentional tasks (Moisala et al., 2016). Similar findings were also reported (Adiguzel et al., 2023) on productivity, engagement, and learning outcomes that may improve as a result of AI's integration, but also suggested that the ethical and practical risks need to be accounted for. The same conclusions were also reached by other researchers (Baido-Anu & Anash, 2023; Firat, 2023).

Another relevant cognitive process is motivation. Specifically referencing motivation in the context of learning, an obvious assumption may be that the presence of AI decreases motivation because learners no longer "have to" complete certain common tasks, such as essays, reflections, analyses, or even writing programming language code because these can be instantaneously generated by ChatGPT. However, this does not appear to be the case

as already research studies agree that the integration of AI in coursework and academic environments greatly increases learner motivation (Ali et al., 2023; Yildiz, 2023).

3.2 Language Processing and Memory

Language processing is an important cognitive function, most rapidly developing within the first couple of years of life. From a neuroscientific perspective, extensive media use is related to decreased brain white matter pathways responsible for reading and language (Hutton et al., 2020), as well as decreased brain connectivity between areas responsible for word recognition, language and cognitive control, and executive brain functioning (Horowitz-Kraus & Hutton, 2018). This was similarly established in other research showing that children with a higher amount of screen time demonstrated significantly poorer language development (Horowitz-Kraus & Hutton, 2018; Duch et al., 2013). However, it is worth noting that children, as more impressionable beings and their brain still being in the development phase, are more prone to these negative effects than adults.

Another cognitive function impacted by technology is memory. Researchers have investigated a commonly asked question, whether easy access to countless information on the internet will affect our semantic memory (i.e., memorizing facts). Their work has demonstrated that our ability to immediately access information online causes the tendency not to remember the information we read because we know it is easily accessible, a phenomenon known as “Digital Amnesia” or “The Google Effect” (Sparrow et al., 2011).

These and other similar important studies were conducted during a time when the internet was becoming more accessible. However, none of these studies reflected the recent proliferation of AI that we are currently subjected to. Hence, the impact of new technology, specifically AI, is presumed to have changed what we have known in this recent period. For example, Sparrow’s findings reflected changes in semantic memory, which only refers to memorizing facts, while other aspects of memory were not affected. However, now we not only have access to information online, but to language models, such as ChatGPT, that can easily write an essay within seconds, or present us with step-by-step instructions for something that we would once have had to figure out by ourselves or by researching multiple information sources. It is, thus, reasonable to postulate that as technology advances, including AI, the same and other aspects of our cognition are bound to be affected in the future as well.

4. Concluding Remarks

Very few aspects of our lives and our society are likely to remain unaffected by AI. Already organizations and governments are working together to maximize benefit for society as it is understood that the future of computing, specifically AI, and correspondingly ethics will be shaped by continued technological advancements, social and cultural acceptances, and regulatory frameworks.

The groundwork for this has already been laid in education. Even though the rate of technology adoption in nearly every aspect of our lives had been on the increase, our access and comfort level in teaching and learning received a boost as a result of the recent pandemic when our physical self suddenly needed to become also virtual. Educational institutions rushed to invest in and integrate as much technology as was feasible to continue their work in a virtual environment. Learners and their instructors managed their digital spaces so well that this type of virtual and independent learning has now become mainstream alongside the traditional classroom mode. Ever since, there has been a debate on how technology will continue to change education. This debate has only intensified now with an ongoing AI movement that will significantly facilitate and give further credence to these developments.

The increased reliance on AI could lead to devaluation of teaching expertise, which can impact training as well as the growth and development of future workforce. Thus, the actual practice of education, although already impacted by automation, must remain human-centric, particularly the activities that deal directly with students. AI will have the important role of augmenting the skills and the work of teachers, ranging from making more efficient lectures and activities, to providing timely feedback and support. It will be important to ensure that AI systems remain subordinate to human command. However, realizing that AI will only continue to develop, attempting to forestall it is simply not possible. Whether we like it or not, AI is going to become an everyday part of our lives, and now is the time to learn to adapt to it.

References

Adıgüzel, T., Kaya, M. H. and Cansu, F. K. (2023) 'Revolutionizing education with AI: Exploring the transformative potential of ChatGPT', Contemporary Educational Technology, 15(3), ep429. doi: 10.30935/cedtech/13152.

Ali, J. K. M. (2022) 'The Impact of Online Learning amid COVID-19 Pandemic on Student Intrinsic Motivation and English Language Improvement', Dirāsāt. Al-ʾulūm Al-insāniyyaẗ Wa-al-iğtimāʾiyyaẗ/DirāSāT. Al-ʿulūM Al-insāNiyyaẗ Wa-al-iğTimāʿiyyaẗ, 49(6), pp. 125–134. doi: 10.35516/hum.v49i6.3726.

Bai, L., Liu, X. and Su, J. (2023) 'ChatGPT: The cognitive effects on learning and memory', Brain-X, 1(3). doi: 10.1002/brx2.30.

Baidoo-Anu, D. and Ansah, L. O. (2023) 'Education in the Era of Generative Artificial Intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning', Social Science Research Network. doi: 10.2139/ssrn.4337484.

Bork, P. (2012) 'How Video games may Enhance Students' learning and cognitive development', The International Journal of Technology Knowledge and Society, 8(1), pp. 43–56. doi: 10.18848/1832-3669/cgp/v08i01/56257.

Buçinca, Z., Malaya, M. B. and Gajos, K. Z. (2021) 'To trust or to think', Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), pp. 1–21. doi: 10.1145/3449287.

Darvesh, N. et al. (2020) 'Exploring the prevalence of gaming disorder and Internet gaming disorder: a rapid scoping review', Systematic Reviews, 9(1). doi: 10.1186/s13643-020-01329-2.

Duch, H. et al. (2013) 'Association of screen time use and language development in Hispanic toddlers', Clinical Pediatrics, 52(9), pp. 857–865. doi: 10.1177/0009922813492881.

Essien, A. et al. (2024) 'The influence of AI text generators on critical thinking skills in UK business schools', Studies in Higher Education, 49(5), pp. 865–882. doi: 10.1080/03075079.2024.2316881.

Firat, M. (2023) 'What ChatGPT means for universities: Perceptions of scholars and students', Journal of Applied Learning and Teaching, 6(1). doi: 10.37074/jalt.2023.6.1.22.

Horowitz-Kraus, T. and Hutton, J. (2017) 'Brain connectivity in children is increased by the time they spend reading books and decreased by the length of exposure to screen-based media', Acta Pædiatrica, 107(4), pp. 685–693. doi: 10.1111/apa.14176.

Hutton, J. et al. (2020) 'Associations between Screen-Based Media Use and Brain White Matter Integrity in Preschool-Aged Children', JAMA Pediatrics, 174(1), e193869. doi: 10.1001/jamapediatrics.2019.3869.

Jia, X.-H. & Tu, J.-C. 2024b. Towards a new conceptual model of AI-Enhanced learning for college students: the roles of artificial intelligence capabilities, general Self-Efficacy, learning motivation, and critical thinking awareness. Systems, 12, 3, pp. 74. URL: https://doi.org/10.3390/systems12030074.

Lacka, E., Wong, T.C. & Haddoud, M.Y. 2021b. Can digital technologies improve students' efficiency? Exploring the role of Virtual Learning Environment and Social Media use in Higher Education. Computers & Education, 163, pp. 104099. URL: https://doi.org/10.1016/j.compedu.2020.104099.

Moisala, M., Salmela, V., Hietajärvi, L., Salo, E., Carlson, S., Salonen, O., Lonka, K., Hakkarainen, K., Salmela-Aro, K. & Alho, K. 2016. Media multitasking is associated with distractibility and increased prefrontal activity in adolescents and young adults. NeuroImage, 134, pp. 113–121. URL: https://doi.org/10.1016/j.neuroimage.2016.04.011.

Ophir, E., Nass, C. & Wagner, A.D. 2009. Cognitive control in media multitaskers. Proceedings of the National Academy of Sciences, 106, 37, pp. 15583–15587. URL: https://doi.org/10.1073/pnas.0903620106.

Yildiz, T.A. 2023. The impact of ChatGPT on language Learners' motivation. Journal of Teacher Education and Lifelong Learning, 5, 2, pp. 582–597. URL: https://doi.org/10.51535/tell.1314355.

What The Phish! Effects of AI on Phishing Attacks and Defense

Shreyas Kumar, Anisha Menezes, Sarthak Giri and Srujan Kotikela
Texas A&M University, College Station, USA

shreyas.kumar@tamu.edu
anisha.menezes@tamu.edu
nb2812@tamu.edu
drsk@tamu.edu

Abstract: The rapid advancement of artificial intelligence (AI) has significantly transformed the landscape of phishing attacks, presenting new challenges for detection and defense. AI-generated phishing emails, which leverage machine learning and natural language processing (NLP), have become increasingly sophisticated, making traditional detection methods ineffective. This research analyzes the evolution and impact of AI-driven phishing attacks, comparing the distinguishing linguistic and contextual patterns of AI-generated versus human-generated phishing emails. The study utilizes a comprehensive dataset, insights from informal discussions with Chief Information Security Officers (CISOs), and an analysis of historical phishing incidents before and after the release of advanced generative models like ChatGPT. Findings reveal that AI-generated phishing emails exhibit higher success rates due to their ability to bypass conventional spam filters and mimic human communication styles. Additionally, the research identifies significant gaps in current defense strategies and recommends a multi-layered security framework that integrates AI-specific detection tools, real-time threat intelligence, and machine learning-based anomaly detection to mitigate these evolving threats. This study emphasizes the need for organizations to proactively adapt to the growing sophistication of AI-powered phishing by implementing advanced defenses that are capable of keeping pace with the rapidly changing cyber threat landscape.

Keywords: AI-generated phishing, Cybersecurity, Machine learning, Data breaches, Defensive strategies

1. Introduction

The rise of artificial intelligence (AI) has significantly heightened the sophistication and effectiveness of phishing attacks, making them increasingly difficult to detect using traditional cybersecurity measures. AI-generated phishing emails utilize advanced natural language processing (NLP) and machine learning techniques to craft highly personalized and contextually relevant messages that can evade conventional spam filters and security systems. These AI-driven attacks pose a serious threat to critical industries like finance, healthcare, and agriculture, where interconnected systems are particularly vulnerable to disruption. For example, a targeted attack on the beef supply chain could manipulate logistics data, disrupt operations, and cause significant economic and public health repercussions. This research analyzes the unique features of AI-generated phishing emails, evaluates current defense mechanisms, and reveals significant gaps in existing security frameworks based on a survey of CISOs. To address these challenges, it proposes a multi-layered security framework incorporating AI-driven threat detection models, real-time threat intelligence, and improved traceability through RFID enhancements. Organizations must act swiftly to implement AI-based defenses, as failure to adapt could lead to severe data breaches, financial losses, and compromised digital infrastructures, underscoring the need for a proactive and integrated cybersecurity approach.

Research Questions

RQ1) How do AI-generated phishing emails compare to traditional phishing emails in terms of detection difficulty and effectiveness in bypassing current email security systems?

RQ2) Is there a growth of AI-based solutions to counter phishing attacks?

2. Background

Phishing attacks, which use deceptive emails or instant messages to trick users into divulging private information, have seen a significant rise in volume and complexity over the past few decades. Initially coined in the mid-1990s, “phishing” described a method used by hackers posing as AOL representatives to steal user credentials (Steves, Greene, and Theofanos, 2019). These early attacks laid the groundwork for what would become one of the most prevalent cyber threats today. As social media gained traction, attackers began exploiting personal information to craft targeted messages in what is now known as spear phishing. The latest evolution in phishing involves the use of artificial intelligence (AI) to generate convincing emails that can easily bypass traditional detection systems (Barrientos, Jacobs, and Dawkins, 2021).

With AI tools like natural language processing and machine learning, cybercriminals can now automate the creation of sophisticated phishing emails at unprecedented speed and scale (Canham, Dawkins, and Jacobs, 2024). This has contributed to a 94% increase in phishing attacks since 2020, with a 27.8% surge in global phishing activities from 2022 to 2023 alone (Proofpoint, 2024; Bolster, 2024). The release of AI tools like ChatGPT in November 2022 has made it easier for attackers to generate AI-phishing emails that are difficult for both humans and current models to detect. Technologically less advanced adversaries, such as smaller nations or organizations with limited resources, that had already been facing distinct difficulties in defending against cyberattacks, now face even greater challenges (Kumar, Nagar, 2024). As a result, cybersecurity defenders must urgently develop and implement AI-specific detection software (Greene, Steves, and Theofanos, 2018).

This research aims to analyze the evolving techniques used in AI-powered phishing and assess the effectiveness of existing countermeasures. By evaluating changes in the accuracy of leading phishing detection software and conducting interviews with Chief Information Security Officers (CISOs), this study provides a comprehensive overview of how organizations are adapting to the rise of AI-phishing and identifies areas for improvement in defensive strategies. The findings will contribute to developing advanced security measures that can effectively combat this growing threat (NIST, 2024; Schwartz, 2022).

3. Related Work

Research in the field of phishing detection and defense has grown significantly in recent years, particularly with the advent of AI-generated phishing attacks. Steves et al. (2019) introduced the "Phish Scale," a novel tool for assessing the difficulty of detecting phishing messages, which has set the foundation for evaluating phishing susceptibility in users. This work was further extended by Barrientos et al. (2021), who developed a framework for scaling the difficulty of detecting AI-generated phishing emails. Canham et al. (2024) expanded on these studies by examining differential phishing susceptibility among various user demographics, highlighting that not all victims react the same way to similar phishing cues. Similarly, Neil et al. (2023) provided an analysis of real-world phishing indicators, revealing how user behavior is influenced by specific email characteristics.

From a machine learning perspective, the National Institute of Standards and Technology (NIST) has contributed significantly to understanding and managing AI risks. In particular, NIST (2024) published a comprehensive taxonomy of adversarial machine learning attacks, detailing common vulnerabilities in AI-based systems and suggesting mitigation strategies. This publication complements the work of Schwartz (2022), which addresses bias in AI systems and its implications for cybersecurity, underscoring the need for balanced datasets and fair algorithmic practices. Additionally, Greene et al. (2018) explored the effectiveness of various cybersecurity strategies against sophisticated phishing attempts, emphasizing the challenges in detecting attacks that exploit advanced natural language processing capabilities.

3.1 Pre AI Attack and Defense

Before the advent of AI, phishing attacks were still on the rise as people became increasingly reliant on the Internet. During the COVID-19 pandemic, many services transitioned to online formats and people used instant messaging and emails as their primary means of communication. This gave attackers greater opportunity to deceive unsuspecting people. In fact, over the course of the pandemic, phishing attacks increased more than 600% (Verizon 2024).

However, defenders were also developing detection models at a rapid pace. Jain & Gupta (2016) suggest using an auto-updated white-list of legitimate sites accessed by the user. When the user tries to visit a website not in the white-list, the browser warns the user not to disclose their sensitive information. This model also checks the legitimacy of a website by extracting the hyperlinks from the source code of the webpage and applying them to a phishing detection algorithm (Jain & Gupta 2016). However, this model only had 86.02% accuracy rate in 2016, and phishing attacks have only grown more advanced since then. Yasin & Abuhasan (2016) also introduced the concept of weighting phishing terms in each email during the analysis phase. Their research tested 5 popular classification algorithms and found that the Random Forest algorithm worked the best with 99.1% accuracy.

More recently, Ho et al. (2019) proposed a large-scale characterization of lateral phishing attacks. Lateral phishing attacks are where the adversary leverages a compromised account to email other users, using the implicit trust between users and information from the compromised account. Ho et al. (2019) developed a classifier that found both identified lateral phishing attacks and those that went undetected. The model had an accuracy rate of 87.3% on their dataset of 87,413,431 emails. Unfortunately, Jain & Gupta's (2016) and Ho et al.'s (2019) models may become obsolete in the face of AI-phishing attacks.

3.2 Attacks Using AI

Phishing emails generated by AI often lack the poor grammar and overly suspicious requests that characterize traditional phishing emails. Additionally, AI-phishing emails can be tailored to each individual recipient and can even mimic the linguistic style of a close acquaintance (Heiding et al. 2024). Many research studies have concluded that the proliferation of AI and machine learning threatens the cybersecurity community (Dash & Sharma 2023; Putra et al. 2024; Vassilev et al. 2024). Heiding et al. (2024) posits that AI-phishing emails have a higher click percentage (30-44%) than traditional phishing emails (19-28%). Additionally, the authors note that "AI significantly increases the incentives to launch phishing attacks by reducing their cost and required revenue." (Heiding et al. 2024). As such, there has been a significant spike in phishing emails since the release of ChatGPT. According to Bolster's *2024 State of Phishing and Online Scam Report*, there's been a 94% increase in phishing attacks since 2020, and there was a 27.8% increase in global phishing activities just from 2022 to 2023 (Bolster 2024). As predicted by Heiding et al. (2024), the reduced costs of AI-phishing have led to an influx of illegitimate emails across the globe. The combination of AI and phishing can be used in scenarios described by Kumar et al. (2024) where a technologically less dependent attacker wants to target a more cyber dependent nation state.

Unfortunately, this increase in AI-phishing emails may have more consequences than just bypassing traditional phishing detection software. Sarno & Neider (2021) analyzed how task factors impacted the accuracy of human identification of phishing emails. The researchers discovered that increasing email load decreased classification accuracy when 50% of the emails were phishing attacks and 50% were legitimate. Participants also had false confidence in their ability to identify phishing emails, and were very "liberal" in their classifications, identifying more emails as legitimate than phishing (Sarno & Neider 2021). Given that Heiding et al. (2024) have observed a spike in phishing emails, AI-phishing emails may not only be hard for humans to classify based on their targeted language; the sheer volume of emails may also decrease the accuracy of human identification.

3.3 Defense Using AI

Given the issues stated above, it is critical that cybersecurity professionals implement AI-phishing specific detection in their products to avoid increased data breaches. Fortunately, some researchers have begun fighting fire with fire by implementing AI in their phishing detection models. Koide et al. (2024) found that LLMs can detect AI-phishing attacks to 98% accuracy. Koide et al.'s (2024) model, called "ChatPhishDetector", utilizes a web crawler to gather information from phishing websites and then generates prompts for LLMs based on the crawled data. Then, the detection model scans for emails that are similar to the responses generated by the LLMs. This system is able to detect multilingual phishing sites by "identifying impersonated brands and social engineering techniques in the context of the entire website," which removes the need to train machine learning models. Similarly, Misra and Rayz (2022) have developed a system that integrates both a multilayer Perceptron (MLP) model and two pretrained Natural Language Processing (NLP) models to detect traditional phishing emails. Misra and Rayz's (2022) model can detect phishing attacks with 97.18% accuracy. Çolhak et al. (2024) also developed a model using NLP, but their research focused on detecting phishing websites rather than phishing emails. Sameen et al. (2020) created a machine learning-based detection system called PhishHaven to identify AI-generated and human-crafted phishing URLs as a response to the deep neural network-based phishing URL generating system called DeepPhish. This model has a 98% success rate when detecting either AI generated or human made phishing urls (Sameen et al 2020). The models discussed here show that ML can be used to detect both AI-generated and traditional phishing attacks to a high degree of accuracy. However, just because these models exist in the literature doesn't mean companies are using AI in their detection models currently. That is what this study aims to evaluate.

4. Methodology

The methodology employed in this study involved three primary approaches: dataset analysis, pilot interviews, and a comparative examination of AI-phishing incidents pre- and post-release of advanced generative models like ChatGPT.

Dataset Analysis: A comprehensive dataset of AI-generated and human-generated phishing emails was collected and analyzed to identify distinguishing linguistic and structural features. Techniques such as Natural Language Processing (NLP) were used to extract patterns and anomalies, which were then compared to conventional phishing indicators. The analysis focused on variations in sentence complexity, tone, and content relevance, aiming to pinpoint attributes unique to AI-generated attacks.

Pilot Interviews: Semi-structured interviews were conducted with nine CISOs across various industries to gain insights into their experiences with AI-driven phishing attacks. The interviews explored perceived levels of

sophistication in AI-phishing emails, current defense strategies, and the effectiveness of AI in their organizations' cybersecurity practices. The qualitative data from these discussions was categorized to highlight common challenges and security gaps.

Comparative Analysis: A historical comparison of AI-phishing incidents before and after the release of advanced models such as ChatGPT was performed. The comparison focused on metrics such as the volume of AI-phishing attempts, success rates, and time-to-detection by traditional systems. This analysis provided a quantifiable measure of how AI models have changed the threat landscape and how defenses need to adapt accordingly.

These combined methods enabled a thorough evaluation of AI-driven phishing threats and provided actionable insights for developing AI-specific defensive strategies.

4.1 Limitations

Human Research approval: The Human research approval takes time and our IRB application is still in process. Hence formal surveys or interviews were not possible. We conducted informal discussions as a pilot study, making sure no personal data was collected and the summary of these discussions to not disclose the identity of the experts we talked to.

Due to ethical reasons, there is a limitation on ChatGPT to create phishing emails. We were limited to small examples of emails, generated by ChatGPT in the context of learning.

5. Findings

Following findings are based on information gathered from journal articles, cybersecurity analysis reports and informal discussions with three Security leaders. For detailed discussions with CISOs we are waiting on IRB approval.

5.1 Attack scenarios

According to the latest State of Phishing assessment report by SlashNext Security (2024), the volume of malicious emails has skyrocketed by a whopping 856% in the last 12 months, with an even more alarming 4151% increase since the launch of ChatGPT in November 2022.

Additionally, the significant differences in Verizon's annual security reports from 2022 and 2024 make the upsurge of phishing attacks and data breaches conspicuous, which can partly be associated with the advancement in AI and introduction of LLMs.

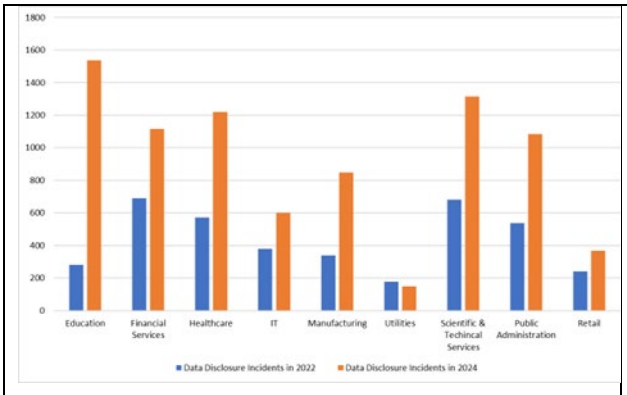


Figure 1: Number of data breaches in major industries between 2022 and 2024

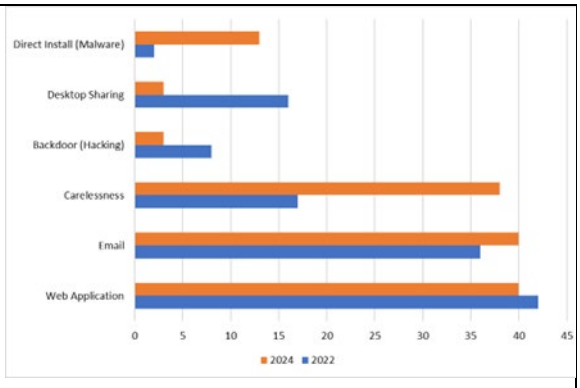


Figure 2: Major Action Vectors in breaches by percentage

(Adapted from Verizon 2022 DBIR and Verizon 2024 DBIR)

Most industries, except Utilities (mining, quarrying, oil extraction etc.) have reported a spike in information leaks, most of them involving system intrusion through hacking and malware. Overall, the percentage of cyber attack incidents covered by Verizon that resulted in data breaches went up from 21.8% in 2022 to 34.9% in 2024, with the contribution of phishing emails rising from 36% in 2022 to about 40% in 2024, highlighting the magnitude of amplification phishing has had when combined with AI.

The biggest phishing attacks from 2022 and 2024 further reinforce the claim. On August 5, 2022, Living Innovations, a global technology firm, fell victim to a social engineering-based phishing attack, one of the top phishing attacks of 2022 (PhishProtection, 2022). Hackers targeted multiple employees by sending phishing emails that tricked them into revealing login credentials. Once inside the network, the attackers gained access to sensitive company information, including intellectual property and operational details.

On August 9, just 4 days later, another phishing data breach occurred when Acorn Financial Services was targeted through an email-based cyber attack, which led to an unauthorized party gaining access to consumer data contained on Acorn's network (JD Supra, 2022). Although both organizations suffered financial losses in millions related to intellectual property compromise, loss of operational data etc., the combined damages paled in comparison to the Pepco Social Engineering attack of February 2024.

The Pepco Group, a major European retailer, faced a phishing attack of unprecedented scale and complexity, reported as a "sophisticated fraudulent phishing attack" by the company. Pepco's breach possibly leveraged AI to craft sophisticated and personalized emails that convincingly targeted the high-level employees. The scope of exposed data was much broader, including not only detailed personal data of the employees and customers but critical infrastructure information as well. As a result, Pepco lost approximately €15.5 million in cash ($17.3 million) and witnessed a decrease in its stock market value (Help Net Security, 2024). The breach was of such extent that it was declared to pose a national security risk. This serves as testament to the fact that AI has enhanced modern phishing into a far more dangerous threat.

5.2 Informal Discussion with CISOs

Based on our pilot study, talking to nine security experts, organizations are currently facing an increased number of sophisticated phishing attacks, particularly Business Email Compromise (BEC), fake login pages, and SMS-based phishing. AI-generated phishing emails are more well-tailored than traditional phishing emails, leveraging personal information such as job titles, trusted contacts, and social media activity to create highly targeted and convincing messages. The respondents also mentioned that AI tends to mimic senior management requests, understand context, and adjust tactics to evade detection.

The emergence of AI has led to the use of deep fakes, advanced social engineering, and more personalized content in phishing campaigns. While the volume of AI-driven phishing emails has not yet significantly increased, it is expected to scale rapidly. The future of AI-driven phishing is likely to involve integration with other cybercrimes, resulting in more damaging and widespread attacks. The automation of phishing campaigns, combined with AI's ability to evade detection, makes these threats more dangerous and accessible to cybercriminals, posing a significant challenge to cybersecurity defenses.

5.3 Defense Scenarios

Organizations are at varying stages of implementing AI to detect and prevent AI-generated phishing attacks. While some have not yet adopted AI in their detection models, there is a strong sense of urgency to do so, given AI's potential to enhance threat detection and response capabilities. For now, most companies are focusing on employee training and awareness campaigns as a first line of defense.

In response to the rise of AI, organizations are encouraging employees to understand its implications and are building internal capabilities. Some have established Centers of Excellence (CoE) for AI and developed in-house tools like AI-based code generation and log parsers. AI-driven defenses are expected to revolutionize cybersecurity by enabling predictive analysis of threats, automated detection and response, and continuous vulnerability assessments. When implemented at scale, these solutions can mitigate risks and minimize damages from AI-powered attacks, providing a proactive approach to securing the organization's digital assets.

In order to evaluate how current email security software has fared against the onslaught of AI-phishing attacks, this paper compared the accuracy rates of the most prominent email security companies from 2022 to 2023. The data was aggregated from the annual security reports by SE Labs, a UK based company that specializes in testing and evaluating cyber security solutions. Since ChatGPT was released in November of 2022 and the SE Labs 2022 report was released in June, this data gives a good snapshot of the height of email security software before AI became mainstream. SE Labs had not yet released their 2024 Email Security Report as of the writing of this paper, so this research uses the 2023 report.

These security reports evaluated how the most popular email security softwares fared against different types of phishing attacks. Additionally, SE Labs calculated the accuracy of each software to detect legitimate emails, to protect against phishing attacks, and then the total accuracy with all factors accounted for.

Table 1: Average Accuracy Rates of the Most Popular Email Security Companies in 2022 and 2023

	2023 Average	2022 Average	2023 Avg.- 2022 Avg.
Total Accuracy	67.2%	77.75%	-10.55%
Protection Accuracy	61%	70.75%	-9.75%
Legitimate Accuracy	93.8%	91.5%	+2.3%
Business Email Compromise (BEC)	49.8%	69%	-19.2%
Social Engineering	31.8%	80.75%	-48.95%
Malware	70.2%	78.75%	-8.55%

(Adapted from SE Labs 2022 Email Security Services Report and SE Labs 2023 Email Security Services Report)

In all categories except for legitimate accuracy, on average, email security software has performed worse from 2022 to 2023. Overall, the total accuracy shows that prominent email security software has decreased in accuracy by 10.55%. When it comes to the percentage of illegitimate emails that these softwares identify (Protection Accuracy), they performed worse against AI-phishing emails than traditional emails. However, these softwares have gotten marginally better at identifying legitimate emails and have a reduced amount of false positives (Legitimate Accuracy).

In terms of the specific types of phishing attacks, the popular softwares perform worse against all versions when it comes to AI-phishing emails. Particularly, they decreased in accuracy by 48.95% on Business Email Compromise (BEC) phishing attacks. This sentiment was echoed in this paper's informal discussions with CISOs, which is discussed further below.

5.3.1 AI offerings for email defense products

We analyzed a range of email security tools, synthesizing data from multiple lists of the most prominent email security tools. Since our aim is not to focus on any particular brand or endorse it, we randomly selected two from an averaged list of top 10 products for deeper analysis: Proofpoint and Mimecast. Mimecast and Proofpoint are both leading cybersecurity companies leveraging artificial intelligence (AI) to enhance protection against sophisticated cyber threats.

5.3.2 case study: Mimecast

Mimecast focuses on AI-powered threat detection and response systems designed to intercept the most sophisticated cyber threats, such as Business Email Compromise (BEC), phishing, and data breaches. The company's approach integrates natural language processing (NLP), social graphing, and computer vision to identify anomalies and detect even the most subtle phishing attempts. By using AI at scale, Mimecast's solutions are able to dynamically learn from new threats and continuously improve the detection accuracy over time. Their AI is layered onto existing security measures, enhancing overall efficacy without relying entirely on automation. This multi-faceted approach ensures a high level of threat protection while maintaining the adaptability needed to counter emerging cyber risks. Mimecast's AI technology is particularly well-suited for organizations looking for comprehensive email and communication security solutions.

5.3.3 Case study: Proofpoint

Proofpoint utilizes AI technologies to protect sensitive data across various platforms, including email, cloud applications, and social media. Their AI models analyze behavioral patterns and contextual signals to detect suspicious activities and prevent data exfiltration. The company emphasizes the use of machine learning

algorithms to identify potential threats before they can cause damage, and also provides real-time monitoring and incident response capabilities. Proofpoint's AI technologies are specifically designed to integrate seamlessly with existing security infrastructures, making them a practical choice for organizations aiming to bolster their defense mechanisms against AI-driven cyber attacks.

Mimecast and Proofpoint are both leveraging AI technologies to provide robust cybersecurity solutions, but their approaches and focus areas vary significantly. Mimecast's AI-driven solutions emphasize comprehensive threat detection for email and communication systems. By using advanced techniques like natural language processing (NLP), social graphing, and machine learning, Mimecast's AI models can detect subtle anomalies that indicate sophisticated phishing attempts or Business Email Compromise (BEC). The company's focus on integrating AI with traditional cybersecurity measures enables a layered security approach, which helps reduce false positives and provides a balanced, adaptive defense against a broad spectrum of cyber threats.

In contrast, Proofpoint's AI technologies focus on protecting sensitive data across multiple platforms, including email, cloud services, and social media. Proofpoint utilizes machine learning algorithms to analyze user behaviors and detect anomalous activities indicative of cyber threats. The company's AI solutions are built to prevent data exfiltration and to provide real-time monitoring, making it particularly strong in protecting against data breaches and insider threats. Proofpoint's integration of AI into existing security infrastructures makes its solutions flexible and easily adaptable for organizations looking to enhance data protection without overhauling their current systems.

In summary, while both Mimecast and Proofpoint use AI to enhance cybersecurity, Mimecast focuses more on threat detection and communication security, whereas Proofpoint specializes in data protection and preventing data loss across various platforms. Both companies highlight that AI should complement existing security measures, providing organizations with a multi-layered defense against evolving cyber threats.

5.4 Informal Discussions with CISOs

The CISOs agreed that AI-driven defense could be very promising; however, multiple warned that such models would require lots of data to work on a large scale. This lack of data seems to be the underlying reason behind why most companies are emphasizing on phishing awareness training programs, while simultaneously trying to implement AI into their current detection softwares.

Organizations are at varying stages of implementing AI to detect and prevent AI-generated phishing attacks. While some have not yet adopted AI in their detection models, there is a strong sense of urgency to do so, given AI's potential to enhance threat detection and response capabilities. For now, most companies are focusing on employee training and awareness campaigns as a first line of defense.

In response to the rise of AI, organizations are encouraging employees to understand its implications and are building internal capabilities. Some have established Centers of Excellence (CoE) for AI and developed in-house tools like AI-based code generation and log parsers. AI-driven defenses are expected to revolutionize cybersecurity by enabling predictive analysis of threats, automated detection and response, and continuous vulnerability assessments. When implemented at scale, these solutions can mitigate risks and minimize damages from AI-powered attacks, providing a proactive approach to securing the organization's digital assets.

Our findings indicate that AI-generated phishing emails are significantly more challenging to detect and are more effective at evading traditional security mechanisms compared to human-generated phishing attempts. And the analysis of security products indicates a trend towards strong AI-based offerings.

6. Future Work

Future research will focus on refining AI-based defense strategies to improve the detection and prevention of AI-generated phishing attacks. This includes developing predictive models that can identify emerging phishing patterns and integrating AI with real-time threat intelligence to enhance response capabilities. Additionally, the use of large language models (LLMs) will be explored to automate the identification of phishing emails, analyze attacker behaviors, and provide adaptive learning systems that can evolve with the threat landscape. Expanding the research to include more diverse datasets and conducting extensive evaluations across different sectors will provide a more comprehensive understanding of AI's role in both offensive and defensive cybersecurity applications. Once we receive the human research approval, we plan to conduct formal and detailed interviews with security experts, practitioners and product designers.

7. Conclusion

The findings of this research indicate that AI-generated phishing and cyber threats are becoming increasingly sophisticated, making traditional cybersecurity defenses insufficient. The study demonstrates that AI-powered attacks can easily evade detection by conventional security mechanisms, thereby escalating the risk of data breaches, financial losses, and organizational disruptions. This evolving threat landscape underscores the necessity for organizations to proactively adapt their cybersecurity strategies to include AI-driven defenses that are capable of identifying, analyzing, and mitigating these advanced threats in real time.

To address this challenge, it is recommended that organizations adopt a multi-layered approach incorporating AI-based threat detection and response systems alongside traditional security measures. Implementing such advanced technologies will enable organizations to more effectively anticipate and counteract AI-enabled attacks. This approach not only enhances detection capabilities but also strengthens overall resilience by enabling rapid adaptation to new and emerging threats.

While integrating AI-driven defenses may require significant initial investment and strategic planning, the long-term benefits in terms of reduced risk and enhanced security far outweigh the costs. Therefore, it is imperative that industry leaders, policymakers, and cybersecurity professionals prioritize the adoption of AI-based solutions to safeguard their digital assets and ensure the stability and integrity of their operations in an increasingly complex cyber environment. A failure to act promptly could leave organizations exposed and vulnerable to the next generation of AI-powered cyber attacks, making immediate action essential to maintaining robust cybersecurity postures.

References

Barrientos, F., Jacobs, J., and Dawkins, S. (2021). S*caling the phish: Advancing the NIST phish scale*. In HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23 (pp. 383-390). Springer International Publishing.

Canham, M., Dawkins, S., and Jacobs, J. (2024, June). *Not All Victims Are Created Equal: Investigating Differential Phishing Susceptibility*. In International Conference on Human-Computer Interaction (pp. 3-21). Cham: Springer Nature Switzerland.

Çolhak, F., Ecevit, M.İ., Uçar, B.E., Creutzburg, R., and Dağ, H. (2024). *Phishing Website Detection through Multi-Model Analysis of HTML Content*. arXiv preprint arXiv:2401.04820.

Dash, B., and Sharma, P. (2023). *Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review*. International Journal of Engineering and Applied Sciences, 10(1).

Dawkins, S., Jacobs, J., and Neil, L. (2023). *Peering into the phish bowl: An analysis of real-world phishing cues*. Indicator, 43, pp.72-88.

Greene, K., Steves, M., and Theofanos, M. (2018). *No phishing beyond this point.* Computer, 51.

Greene, K.K., Tamborello, F.P. (2017). *Exploratory lens model of decision-making in a potential phishing attack scenario*. US Department of Commerce, National Institute of Standards and Technology.

Heiding, F. et al. (2024). *Devising and detecting phishing emails using large language models*, IEEE Access, 12. doi:10.1109/access.2024.3375882.

Ho, G., Cidon, A., Gavish, L., Schweighauser, M., Paxson, V., Savage, S., Voelker, G.M., and Wagner, D. (2019). *Detecting and characterizing lateral phishing at scale*. In 28th USENIX security symposium (USENIX security 19).

Jain, A.K., and Gupta, B.B. (2016). *A novel approach to protect against phishing attacks at client side using auto-updated white-list*, EURASIP Journal on Information Security, 2016(1). doi:10.1186/s13635-016-0034-3.

Koide, T., Fukushi, N., Nakano, H., and Chiba, D. (2023). *Detecting phishing sites using chatgpt*. arXiv preprint arXiv:2306.05816.

Kumar, S., Kocian, L. and Loo, L. 2024. *Blockchain Applications for Cyber Liability Insurance*. 2nd International Conference on Blockchain, Cybersecurity and Internet of Things, BCYIoT 2024.

Kumar, S., and Nagar, G. (2024). *Threat Modeling for Cyber Warfare against Less Cyber-dependent Adversaries*. 23rd European Conference on Cyber Warfare and Security (ECCWS), 27-28 June, Jyväskylä, Finland.

Mimecast (2024). *AI-powered security organizations can trust*, *Ai Cybersecurity*. Available at: https://www.mimecast.com/ai-cybersecurity/ (Accessed: 29 September 2024).

Misra, K. and Rayz, J.T. (2022). *LMS go phishing: Adapting pre-trained language models to detect phishing emails*. 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 3, pp. 135–142. doi:10.1109/wi-iat55865.2022.00028.

Proofpoint (2023) *Proofpoint AI technologies that protect sensitive data*, *Proofpoint US*. Available at: https://www.proofpoint.com/us/resources/solution-briefs/proofpoint-ai-technologies-protect-sensitive-data (Accessed: 29 September 2024).

Putra, F.P.E., Zulfikri, A., Arifin, G., and Ilhamsyah, R.M. (2024). *Analysis of Phishing Attack Trends, Impacts and Prevention Methods: Literature Study*. Brilliance: Research of Artificial Intelligence, 4(1), pp.413-421.

Sameen, M., Han, K., and Hwang, S.O. (2020). *Phishhaven—an efficient real-time AI phishing urls detection system*, IEEE Access, 8. doi:10.1109/access.2020.2991403.

Sarno, D.M., and Neider, M.B. (2021). *So many phish, so little time: Exploring email task factors and phishing susceptibility*, Human Factors: The Journal of the Human Factors and Ergonomics Society, 64(8), pp. 1379–1403. doi:10.1177/0018720821999174.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., and Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (Vol. 3, p. 00). US Department of Commerce, National Institute of Standards and Technology.

Steves, M.P., Greene, K.K., and Theofanos, M.F. (2019). *A phish scale: rating human phishing message detection difficulty*. In Workshop on usable security (USEC).

Vassilev, A., Oprea, A., Fordyce, A., and Anderson, H. (2024). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (No. NIST Artificial Intelligence (AI) 100-2 E2023). National Institute of Standards and Technology.

Yasin, A., and Abuhasan, A. (2016). *An intelligent classification model for phishing email detection*. arXiv preprint arXiv:1608.02196.

Verizon (2022) *2022 Data Breach Statistics by industry*, *Verizon Business*. Available at: https://www.verizon.com/business/resources/reports/dbir/2022/intro-to-industries/ (Accessed: 27 September 2024).

Verizon (2024) *DBIR report 2024: Industries - introduction*, *Verizon Business*. Available at: https://www.verizon.com/business/resources/reports/dbir/2024/industries-intro/ (Accessed: 27 September 2024).

Governance Considerations of Adversarial Attacks on AI Systems

Nombulelo Faith Lekota
CTS-Fundação Getulio Vargas (FGV), Rio de Janeiro, Brazil

Nombu30@gmail.com

Abstract: Artificial intelligence (AI) is increasingly integrated into various aspects of daily life, but its susceptibility to adversarial attacks poses significant governance challenges. This paper explores the nature of these attacks, where malicious actors manipulate input data to deceive AI algorithms and their profound implications for individuals and society. Adversarial attacks can undermine critical AI applications, such as facial recognition and natural language processing, leading to privacy violations, biased outcomes, and eroding public trust. The discussion emphasizes understanding the threat vectors associated with adversarial attacks and their potential repercussions. It advocates for robust governance frameworks encompassing risk management, oversight, and legislative measures to protect AI systems. Such frameworks should prioritize AI technologies' confidentiality, integrity, and availability (CIA) while ensuring compliance with ethical standards. Furthermore, the paper examines various strategies for mitigating risks associated with adversarial attacks, including training and continuous monitoring of AI systems. It highlights the importance of accountability among developers and researchers in implementing preventive measures that align with principles of transparency and fairness. Organizations can enhance security and foster public trust by integrating legislative frameworks into AI development standards. As AI technologies evolve, continuous review of governance practices is essential to address emerging threats effectively. This paper ultimately focuses on the critical role of comprehensive governance in safeguarding AI systems against adversarial attacks, ensuring that technological advancements benefit society while minimizing risks.

Keywords: Artificial intelligence (AI), Adversarial attacks, Governance, Frameworks, Confidentiality, Integrity, Availability (CIA)

1. Introduction

Today, artificial intelligence (AI) significantly influences people's lives and addressing governance issues related to adversarial attacks on AI systems is essential. Understanding the impact of these attacks and taking proactive measures to mitigate the risks is crucial. Adversarial attacks involve intentionally manipulating data to deceive AI algorithms, creating significant challenges for governance. AI systems, essential for various technological applications such as facial recognition and language understanding, are vulnerable to these attacks, potentially affecting individuals and society profoundly.

Governance of AI systems vulnerable to adversarial attacks is complex and crucial. This paper outlines the following sections: Section 2 covers adversarial attacks on AI models, Section 3 discusses their impacts, Section 4 addresses associated risks and challenges, Section 5 explores governance frameworks for securing AI systems, and Section 6 offers recommendations. Finally, Section 7 concludes the paper.

The following section outlines different types of adversarial attacks in AI systems.

2. Understanding Adversarial Attacks on AI Systems

Artificial intelligence (AI) technologies hold significant potential to revolutionize society and people's lives, spanning commerce, health, transportation, cybersecurity, and environmental sustainability. The technologies generate outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives. However, they also pose risks and challenges that could adversely affect individuals, groups, organizations, communities, society, and the environment (Australian Institute of Company Directors & UTS Human Technology Institute, 2024; Tabassi, 2023a).

Adversarial attacks target vulnerabilities in AI systems by subtly manipulating input data, resulting in incorrect outputs or altered behaviour. Attackers may render data inaccessible, preventing authorized users from accessing it. They can also compromise data integrity by altering or corrupting information, undermining its accuracy and trustworthiness. Additionally, unauthorized access to sensitive data can lead to privacy violations, exposing confidential information to unintended parties (Vassilev et al., 2024).

Adversarial attacks are deliberate methods of manipulating the AI system input data to mislead the system and produce inaccurate output data (Souza, 2023). Such attacks present significant security breaches and privacy violations, posing challenges in detecting and mitigating current and future attacks. Adversarial attacks can trick a target model into making completely inaccurate predictions by modifying original images (Deng et al., 2020).

The paragraphs below present the adversarial attacks, vulnerabilities of AI models, and their intended effects.

A *prompt injection* is a security flaw that manipulates large language models (LLMs) such as ChatGPT and Google Bard using engineered malicious prompts and affecting prompt-based learning. The attack can manipulate the language model's output via engineered malicious prompts (Liu et al., 2023). The attack enables malicious actors to manipulate the prompts provided to the AI models to bypass the built-in restrictions to cause unintended actions (OWASP, 2023). The consequences of prompt injection vulnerabilities are serious, as hackers can manipulate AI models to give instructions for illegal activities. The result can be revealing API keys and secrets, compromising system and data security. AI developers must address this risk during system development.

After developing AI systems, the model outputs must be correctly validated, sanitized, or managed before application use. However, *insecure output handling* is a vulnerability that affects insufficiently validated and sanitized LLMs. The effect can be passed downstream to other components and systems and influence the content generated by these models, providing incorrect information (OWASP, 2023). In developing and deploying AI, paying close attention to security practices is crucial to mitigate potential vulnerabilities and attacks caused by insecure output handling.

Data poisoning is a cyber-attack where a malicious actor manipulates the training data sets. The attackers intentionally skew the results to sabotage the system operations and influence the model's predictions and decision-making capabilities (Ballejos, 2024). The manipulated models provide false, misleading, or malicious data during training. This challenge requires developers' attention, especially when developing AI systems.

A *Model Denial of Service (Model DoS)* attack is a vulnerability that allows an attacker to use up many resources from the AI system. An attacker overloads an LLM with resource-intensive interactions, reducing service quality and increasing costs. Manipulating the LLM's context window is also a significant security concern (OWASP, 2023). This complex architecture issue requires a secure ICT infrastructure.

On the other hand, *supply chain vulnerabilities* affect different components and dependencies of AI applications. The AI system vulnerabilities affect the accuracy of training data models and deployment platforms. The attack causes biased outcomes, security breaches, and system failures. *Sensitive information disclosure* attacks occur when confidential data is unintentionally released due to insufficient security measures. When applications fail to protect sensitive information adequately, it can lead to unintended disclosure to unauthorized parties, with severe implications for individuals and organizations (OWASP, 2023).

Inadequate security measures to secure LLM plugins can cause an *insecure plugin design* vulnerability. The attackers produce malicious requests that could result in undesired outcomes, such as remote code execution. The vulnerability relates to how developers design programs, architect solutions, and implement security practices. The other significant development error is called *excessive agency***,** where LLMs misconfigure permissions and have excessive functionality. Lack of model security permission oversight can contribute to excessive agency (OWASP, 2023).

It is important to note that *overreliance* on Large Language Models (LLMs) without proper oversight or validation can result in security breaches, misinformation, miscommunication, legal problems, and damage to one's reputation. The effect can lead to inaccurate information, security weaknesses, and unintended consequences. In addition, *model theft* vulnerability refers to a cyber-attack that allows hackers to access and duplicate model data. Insecure AI plugins can facilitate this, presenting a significant development challenge.

A significant AI development vulnerability is a *black-box scenario attack*. While attackers cannot access the model's internal parameters or architecture, they can interact with it through its outputs, using trial and error to discover vulnerabilities and craft adequate adversarial inputs (OWASP, 2023).

The section discussed the types of adversarial attacks on AI systems and their significant impact. The following section presents real-world scenarios of adversarial attacks.

3. Adversarial Attacks on AI Systems: Impacts and Implications

In this section, it is essential to highlight specific scenarios that demonstrate the potential severity, weaknesses, and flaws in AI systems. Presenting these examples conveys the gravity of the situation and highlights the need for robust security measures to mitigate these risks. These scenarios will help stakeholders grasp the real-world implications of the AI system's vulnerabilities and recognize the importance of addressing them proactively.

Adversarial attacks on autonomous vehicles pose a serious threat to road safety by causing misinterpretations of road signs, lane markings, and the presence of other vehicles. Such attacks can manipulate sensor inputs like lidar and radar, disrupting the vehicle's perception capabilities. This vulnerability increases the risk of accidents and compromises the safety and security of both the vehicle and its occupants (Trent, 2024).

Medical imaging systems can be targeted by attacks where malicious actors alter images to deceive AI, leading to misdiagnoses and significant risks to patient safety. For example, an attacker might modify an X-ray to make a healthy bone look fractured or manipulate a CT scan to introduce false abnormalities. Such changes can result in incorrect treatment plans, unnecessary procedures, and delayed diagnoses, ultimately compromising patient health (Trent, 2024).

The researchers at Carnegie Mellon University demonstrated that specially designed glass frames can deceive even the most advanced facial recognition software. These glasses make the wearer nearly invisible to these automated systems and trick them into identifying the wearer as someone else. The researchers could assume different identities by adjusting the patterns printed on the glasses (Vincent, 2016).

Internet trolls manipulated Microsoft's AI chatbot, Tay, into generating offensive content, resulting in its prompt shutdown. This example highlights the potential for adversaries to exploit vulnerabilities in AI systems to produce harmful outputs. These real-world examples underscore the need for robust defences against adversarial attacks. The consequences of the attacks pose risks and challenges related to AI adversarial attacks, which will be discussed next. The attacks can have a significant impact on individuals, organizations, and society as a whole.

4. Risks and Challenges Associated with Adversarial Attacks

The preceding sections examined the nature of adversarial attacks on AI, real-world scenarios, and their effects on operational performance. This section focuses on assessing the potential risks and challenges these attacks pose and the extent of the damage they can inflict (Bai et al., 2021). First, the risks are presented, followed by the challenges.

4.1 Risks of Adversarial Attacks

AI systems are developed to analyze data, detect patterns, and make predictions or decisions based on that information. However, adversarial attacks exploit vulnerabilities in AI algorithms by leveraging their reliance on specific patterns or features. By subtly manipulating the input data, adversaries can deceive the AI system into producing inaccurate or unexpected results. This presents substantial risks in various domains where the dependability and precision of AI-driven decisions are crucial (Trent, 2024).

Adversarial attacks on AI systems present substantial risks that can lead to dire consequences, including security vulnerabilities such as data breaches and disseminating harmful content. Furthermore, the reputation of organizations utilizing compromised AI systems is at significant risk. In the financial sector, adversaries can manipulate AI-driven fraud detection models through data poisoning, which causes the systems to classify fraudulent transactions as legitimate incorrectly. This malicious interference can lead to considerable financial losses, customer trust, reputation, and financial health.

Similar vulnerabilities exist within supply chains, where the integrity of training data, machine learning models, and deployment platforms can be compromised. Such vulnerabilities may result in biased outcomes, security breaches, or complete system failures. The risk of sensitive information disclosure manifests in various forms, including the exposure of session tokens, passwords, credit card details, and other confidential data.

Insecure design practices can create vulnerabilities that compromise organizational security and expose sensitive data to unauthorized access. Granting large language models (LLMs) excessive permissions may lead to harmful actions, such as executing unauthorized commands or disclosing confidential information. Additionally, over-reliance on LLMs without proper validation can produce incorrect or unsafe content, known as hallucination or confabulation.

Adversarial attacks on AI systems pose serious risks, including financial loss, reputational harm, and security breaches. Thus, the need for effective risk management strategies is required. Organizations should promote a risk management culture throughout AI system development. The NIST AI Risk Management Framework (Tabassi, 2023b) highlights the importance of processes and documentation to identify and manage AI system risks. This involves aligning risk management with organizational policies and addressing technical and legal considerations related to third-party systems and data throughout the product lifecycle (Tabassi, 2023b).

4.2 Challenges Imposed by Adversarial Attacks

The vulnerabilities of AI systems and the dynamic nature of attack strategies contribute to the difficulty in safeguarding AI systems. Detecting and preventing adversarial attacks on AI systems poses significant challenges for developers and security professionals. AI systems have complex architectures that make them vulnerable to cyber threats. Varying levels of threat actors can infiltrate these systems, and the lack of standard cybersecurity frameworks complicates data exchange. Protecting classified information and addressing underreporting and insecure communication channels is vital to enhance data security.

Education is vital for developers and users to understand AI cyber threats. Information sharing among stakeholders is essential for preparedness during incidents. As AI attacks evolve, a robust cybersecurity governance framework is needed for effective collaboration, detection, response, and recovery. Formulating and implementing risk mitigation plans is critical to overcoming cybersecurity response challenges (Lekota & Coetzee, 2021).

The following section addresses the need for practical solutions to adversarial attacks in AI development.

5. Governance Frameworks for Secure AI Systems

Governance refers to the decision-making structures and processes organizations create to manage and secure AI systems. Progress has been made in developing guidelines to support organizations in ensuring their AI system remains secure. This section presents some of the developed AI security governance standards and practices.

AI security frameworks, such as the NIST AI Risk Management Framework (Tabassi, 2023a), Open Worldwide Application Security Project (OWASP) Top 10 for Large Language Models (LLMs) (OWASP, 2023), Google's Security AI Framework (Google, n.d.), and the NCSC Guidelines for secure AI system development (NCSC, n.d.), provide a foundation for stakeholders to develop secure and responsible AI systems. The standards and frameworks are widely used; however, this area needs more research. While organizations are wrestling with technological advancements and emerging adversarial attacks, the standards and framework underscore the importance of responsible development and governance processes (Vassilev et al., 2024).

Aligning with established guidelines from various organizations is essential for effective AI security governance. Critical contributions to AI security governance come from governments, international organizations, and technology companies. The developed frameworks can facilitate multinational agreements, collaboration, and standard AI cybersecurity mechanisms. The following paragraphs highlight some critical frameworks and standards that can guide the recommendations presented in this paper.

The universal cyber governance model (Sarri et al., 2023) emphasizes stakeholder participation, transparency, and accountability, encompassing strategic, political, technological, and operational categories. Organizations are encouraged to create a comprehensive AI cybersecurity strategy with objectives, an operational framework, funding models, and AI incident response teams. Political governance clarifies roles to enhance cooperation, while technical governance aligns AI systems with secure development standards. Operational governance focuses on building a skilled workforce aware of cyber threats and promotes responsible AI use. The model can be adapted as an effective governance model for developing secure AI systems.

Organizations need robust governance frameworks encompassing risk management, interdisciplinary collaboration, transparency, and accountability to address adversarial attack risks. This involves ensuring security throughout the AI lifecycle with regular vulnerability assessments and tailored security measures. Collaboration among AI developers, cybersecurity experts, and legal professionals enhances vulnerability understanding and supports comprehensive security protocols. Additionally, prioritizing transparency in AI operations helps stakeholders understand decision-making processes, which can identify potential vulnerabilities and ensure accountability during attacks (Musser et al., 2023).

As adversarial AI threats evolve, so must the legal frameworks governing AI systems. Clear guidelines on accountability for AI-related incidents can help mitigate risks and enhance public trust in AI technologies (Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS), n.d.; Musser et al., 2023). Regarding AI governance, it is crucial to have strategies focusing on data sanitization, continuous monitoring, model training, and adaptation to prevent adversarial attacks effectively. These strategies involve regularly ensuring the accuracy and validity of training data to avoid the introduction of harmful inputs that could compromise the reliability of AI models.

Proactive adversarial training approaches are essential as they help models become more resistant to attacks and learn to identify and withstand the manipulation of input data. Additionally, organizations should set up continuous monitoring systems to detect any unusual patterns in AI behaviour that may signal an ongoing attack and adjust models to address new adversarial threats (Musser et al., 2023; Paloalto, n.d.; Tang, 2024).

The challenges of adversarial attacks in AI development require practical solutions and robust governance frameworks. Existing AI security frameworks are vital for responsible development, and collaboration among organizations is essential for effective governance. An effective governance model should emphasize stakeholder participation, transparency, and accountability, incorporating risk management and legal principles. Clear accountability guidelines for AI incidents are crucial for mitigating risks and building public trust. Organizations must prioritize data sanitization, continuous monitoring, and model training to prevent adversarial attacks effectively.

The following section presents integrated cybersecurity governance that can guide organizations to address risks and challenges for secure AI systems.

6. Recommendations and Guidelines

This section recommends a model to solve the challenges and risks associated with AI discussed in the previous sections. Figure 1 illustrates an integrated AI security governance model that organizations can consider when addressing adversarial attacks on AI systems. The model comprises seven components: the risk management governance adapted from the NIST AI Risk Management Framework (Tabassi, 2023a), leadership and oversight, development and deployment, AI regulations, user awareness, incident response, and continuous monitoring highlighted in the diagram and explained below.



Figure 1: AI security governance model

Organizations can implement AI risk management governance, involving six critical practices, to exploit AI's potential while reducing adversarial risks. Essential components include establishing clear policies, processes, and practices for mapping and managing AI risks. This framework should encompass reliable AI principles, transparent risk management, and accountability structures to empower teams. Diversity and inclusivity within AI teams are vital for well-rounded decision-making and effective risk management.

Organizations must promote a culture of critical thinking and safety in every AI development and deployment phase, ensuring that risks and impacts are documented and communicated. Engaging diverse AI stakeholders

for feedback on potential risks is also crucial. Organizations should develop comprehensive policies to address third-party software and supply chain risks, establishing contingency plans when necessary. Leadership should create robust internal governance mechanisms integrating security policies throughout the AI ecosystem.

Security should be prioritized from the development stage, incorporating safeguards such as encryption and strict access controls to protect sensitive data. Aligning AI initiatives with evolving regulations fosters transparency and accountability. Regular evaluations for vulnerabilities and continuous monitoring of AI systems are essential for maintaining security. Incorporating human oversight in critical operations offers an additional layer of risk mitigation.

Organizations should adopt a comprehensive AI risk management approach. This includes establishing accountability, promoting diversity in decision-making, and fostering open communication about AI risks. Engaging stakeholders in system design, addressing third-party software and data risks, and having contingency plans for high-risk incidents are essential practices. By integrating these elements into an AI security framework, organizations can enhance the resilience and reliability of their AI systems against emerging threats.

7. Conclusion

While AI offers significant benefits, the potential for adversarial attacks requires a robust approach, balancing operational efficiency with adequate safeguards. Protecting AI systems is complex due to their interconnected architecture and evolving cyber threats. Additionally, the lack of standardized cybersecurity frameworks and diverse skill sets among threat actors complicate security efforts.

Educating developers and users about AI-related cyber threats and fostering collaboration among stakeholders is essential for effective incident management. Governance of AI systems in the face of these challenges requires comprehensive frameworks that integrate risk management, collaboration, transparency, and effective mitigation strategies. As AI evolves, so must the governance practices to ensure security and effectiveness. In our fast-changing digital landscape, proactive AI governance is crucial for mitigating risks. Ongoing research and collaboration are needed to advance AI security. By using real-world examples and practical recommendations, organizations can enhance the resilience of their AI systems, safeguarding their integrity and reliability against a dynamic threat landscape.

Acknowledgements

I want to express my heartfelt gratitude to Prof. Luca Belli, Coordinator of the CTS-FGV and CyberBRICS, and Larissa Galdino de Magalhães Santos, PhD, from the CyberBRICS Research Program, for their invaluable support and guidance throughout the process of writing my paper. I would also like to thank Dan Remenyi, PhD, the Director of Academic Conferences & Publishing International and Academic Bookshop, for the opportunity to contribute to ICAIR 2024. Your insights and encouragement have been instrumental in shaping my research and enhancing the quality of my work.

References

Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS). (n.d.). *Artificial Intelligence Risk & Governance*. AI at Wharton. Retrieved August 30, 2024, from https://ai.wharton.upenn.edu/white-paper/artificial-intelligence-risk-governance/

Australian Institute of Company Directors, & UTS Human Technology Institute. (2024). *A Director's Guide to AI Governance*. policy@aicd.com.au

Bai, T., Zhao, J., Zhu, J., Han, S., Chen, J., Li, B., & Kot, A. (2021). AI-GAN: ATTACK-INSPIRED GENERATION OF ADVERSARIAL EXAMPLES. *Proceedings - International Conference on Image Processing, ICIP*, *2021-September*, pp. 2543–2547. https://doi.org/10.1109/ICIP42928.2021.9506278

Ballejos, L. (2024, May 28). *Data Poisoning: The Newest Threat in Artificial Intelligence and Machine Learning*. NinjaOne. Retrieved August 28, 2024, from https://www.ninjaone.com/blog/data-poisoning/

Deng, Y., Zheng, X., Chen, C., Lou, G., & Kim, M. (2020). An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models. *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10.

Google. (n.d.). *Google's Secure AI Framework (SAIF)*. Google. Retrieved August 30, 2024, from https://safety.google/cybersecurity-advancements/saif/

Lekota, F., & Coetzee, M. (2021). Aviation Sector Computer Security Incident Response Teams: Guidelines and Best Practice. *20th European Conference on Cyber Warfare and Security*, 1–12. https://doi.org/10.34190/EWS.21.028

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2023). *Prompt Injection attack against LLM-integrated Applications*. http://arxiv.org/abs/2306.05499

Musser, M., Spring, J., Liaghati, C., Rohrer, D., Elliott, J., Chowdhury, R., Lohn, A., Shankar, R., Kumar, S., Martinez, C., Frase, H., Rodriguez, M., Hermanek, S., Dempsey, J. X., Leong, B., Grant, C. D., & Bansemer, J. (2023). *Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications*. https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/

NCSC. (n.d.). *Guidelines for secure AI system development*. National Cyber Security Centre. Retrieved August 30, 2024, from https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

OWASP. (2023). *OWASP Top 10 for LLM*. Retrieved July 30, 2024, from https://owasp.org/www-project-top-10-for-large-language-model-applications

Paloalto. (n.d.). *What Is Adversarial AI in Machine Learning?* Paloalto Networks. Retrieved August 30, 2024, from https://www.paloaltonetworks.com/cyberpedia/what-are-adversarial-attacks-on-AI-Machine-Learning

Sarri, Anna., Fernández Bascuñana, Gema., Gross, A.-Kristin., Chiarelli, Federico., Preasca, Marina., & European Union Agency for Cybersecurity. (2023). *Building effective governance frameworks for the implementation of national cybersecurity strategies.*

Souza, C. (2023, June). *Adversarial Attacks on AI/ML Models: Everything You Need to Know*. LinkedIn. Retrieved August 30, 2024, from https://www.linkedin.com/pulse/adversarial-attacks-aiml-models-everything-you-need-know-d-souza/

Tabassi, E. (2023a). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

Tabassi, E. (2023b). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

Tang, O. (2024, April). *How to attack (and defend) an AI system: a primer on security*. Clifford Chance. Retrieved July 30, 2024, from https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2024/04/how-to-attack-and-defend-an-ai-system-a-primer-on-security.html

Trent, R. (2024, February 29). *Adversarial Examples in AI Addressing the risks of AI systems being fooled by carefully-crafted inputs*. Blog. https://rodtrent.substack.com/p/adversarial-examples-in-ai

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). *Adversarial machine learning: A Taxonomy and Terminology of Attacks and Mitigations*. https://doi.org/10.6028/NIST.AI.100-2e2023

Vincent, J. (2016, November 3). *These glasses trick facial recognition software into thinking you're someone else*. The Verge. https://www.theverge.com/2016/11/3/13507542/facial-recognition-glasses-trick-impersonate-fool

Mapping Software-Engineering Industry AI Use to Software-Engineering Curriculum: Developing the AI-USE Framework

Addison Lilholt and Thomas Heverin
The Baldwin School, Bryn Mawr, USA

alilholt@baldwinschool.org
thomas.heverin@baldwinschool.org

Abstract: Estimates predict a global deficit of 4 million software engineers by 2025, further complicated by the software engineering (SE) industry's escalating use of artificial intelligence (AI). To tackle this issue, our research suggests that computer science (CS) curricula in middle and high schools need to be updated to incorporate SE industry segments that significantly employ AI. This strategic curriculum alignment is significant for preparing a workforce equipped to meet future industry demands. Our initial analysis involved reviewing nine international AI education guidelines to evaluate current methods for integrating AI into SE education. The findings indicated a pronounced lack of specific guidance connecting AI applications in the SE industry with educational content. To address this, we performed a systematic literature review of 12 research papers focusing on AI's role across the SE industry, followed by multiple rounds of inductive content analysis. An industry segment was deemed "essential" if it was referenced in 75% or more of the papers' findings. Through this method, we identified 10 essential SE industry segments for inclusion in CS education: software development, software maintenance, process improvement, software economics, knowledge management, project management, software testing, software security, quality assurance, and deployment and operations (DevOps). These findings led to the creation of the AI-USE (Artificial Intelligence Usage in Software Engineering) framework, which maps these 10 key segments to the predominant uses of SE in the industry as identified in the literature. Further inductive content analysis helped us develop subsegments for these essential areas. Ongoing framework development involves refining these subsegments and gathering feedback from industry and academic professionals. We anticipate that the fully developed AI-USE framework will significantly enhance SE education, equipping the next generation of software engineers with the AI proficiency required to address the industry's evolving demands.

Keywords: Artificial intelligence in education, Generative AI, Computer science education, Software engineering, SE education

1. Introduction

The demand for software engineers represents an immediate concern especially since the software engineering (SE) industry is expected to grow 25% in the next four years (Stavaridis, 2023). This growth exacerbates the challenges in "...developing complex software systems that meet quality standards and deadlines, as traditional methods involve time-consuming and repetitive tasks often resulting in delays and cost overruns" (Stavaridis, 2023). The complexity of modern software systems necessitates more efficient and effective development methods.

Artificial intelligence (AI) tools like GitHub Copilot and ChatGPT are transforming the SE industry. The increasing reliance of the SE industry on AI is transforming development practices across all stages of the software development life cycle (SDLC). AI tools support human developers by automating routine tasks, enabling them to focus on more creative and complex aspects of software engineering. Additionally, AI tools can provide the ability to "...automatically document software codes and provide debugging routines that offer automated suggestions for improvement or deliver code examples for particular problem sets" (Barenkamp, 2021). Additionally, AI tools can increase efficiency and productivity by using neural networks that can "...assist software coding, allowing for natural language to be processed into software code" will increase efficiency and productivity (Stavaridis, 2023).

Although AI technologies can provide benefits to software engineers, the SE industry still faces challenges in ensuring that the workforce has the necessary skills to leverage these AI technologies effectively. Focusing on the education of future software engineers can help mitigate these challenges. As AI-driven tools like GitHub Copilot and automated testing frameworks become more prevalent, they introduce new security vulnerabilities and testing challenges that must be addressed to ensure the safety and reliability of software systems. These challenges are especially key in the context of the software supply chain, where the rapid spread of AI-generated or AI-tested code can exacerbate security risks on a global scale. Addressing these issues within SE education is essential to prepare the next generation of engineers for the complexities of AI-integrated development environments. Aligning middle and high school computer science (CS) curricula with industry needs in software engineering and computer science, particularly in the area of AI, is essential for several reasons. First, it ensures that students acquire relevant skills that are in high demand in the job

market, improving their employability. The rapid advancement and integration of AI in various industries necessitate a workforce proficient in AI technologies and methodologies. By incorporating AI experience into early CS education, students can develop a strong foundation that prepares them for future careers in these emerging fields. Second, early exposure to AI concepts and tools can stimulate students' interest and engagement in computer science. This engagement can foster a deeper understanding and appreciation of the subject, potentially leading to increased enrolment in CS courses and a more diverse and robust pipeline of future software engineers.

Incorporating SE industry-relevant skills and knowledge into middle and high school curriculum can help bridge the gap between education and practical application. Garousi et al, 2019 stated that "...it is important to synthesize and aggregate the results of all of these studies to provide a single "consolidated" view on how to align SE education with industry needs, and to find out the most important skills in the industry and the knowledge gaps, i.e., the SE topics that should get more education and training." This alignment ensures that the education system produces graduates who are ready to meet the current and future needs of the industry, contributing to the overall growth and innovation within the field. Garousi et al. (2019) conducted a review that showed "...the importance of SE professional practice and soft skills in general, the importance of certain areas in SE education (especially requirements, design and testing), knowledge gaps in specific areas of SE (especially configuration management, SE models and methods as well as SE process), and the importance of real- world examples in SE courses." However, their research did not focus on the integration of AI in SE industry practice.

As mentioned previously, AI tools like GitHub Copilot are transforming the software development process by automating routine tasks and providing contextually relevant code suggestions, which can significantly improve efficiency and productivity. However, the successful integration of these technologies in professional environments requires a workforce that is adept at using and managing AI tools, highlighting the importance of integrating these experiences into educational curricula while maintaining a focus on strong oversight and best practices in security. The overarching objective of our research consists of developing the AI-USE (Artificial Intelligence Usage in Software Engineering) framework to give guidance to SE educators in effectively conveying methods of AI usage in the SE industry, better preparing SE students for the workforce.

2. Methodology

The development of the AI-USE framework consisted of multiple steps including: reviewing current AI education guidelines, conducting a systematic review of research papers, identifying industry segments, and identifying industry subsegments. The authors of the paper both have over 10 years of CS teaching experience as well as several years of software development experience. Both authors have presented on the use of AI in the high school and middle school levels at cybersecurity conferences and AI symposiums. Additionally, one author has conducted qualitative and quantitative research for over 10 years and holds a U.S. patent for a SE project that was completed for the U.S. Department of Defense (DoD).

2.1 AI Guidelines Review

For our framework development process we began by examining nine international AI education guidelines published within the last three years. We conducted searches to identify AI education guidelines and focused on ones that addressed AI in education from 2021 to present. Table 1 shows the nine education guidelines selected and reviewed for the current study. The main focus points of the educational guidelines included AI literacy and comprehensive education, ethical AI utilization, professional development and competency, AI in curriculum design and other areas.

Table 1: Nine International AI Education Guidelines

Title (Organization)	Year
Computing Curricula 2020 CC2020 Paradigms for Global Computing Education (Association for Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE))	2021
Artificial Intelligence and Education: A critical view through the lens of human rights, democracy, and the rule of law (Council of Europe)	2022
Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world (Springer)	2023

Title (Organization)	Year
Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations (U.S. Department of Education)	2023
AI literacy in K-12: a systematic literature review (International Journal of STEM Education)	2023
Artificial intelligence (AI) learning tools in K-12 education: A scoping review (Journal of Computers in Education)	2024
Guidance, Considerations, & Intentions for the Use of Artificial Intelligence in West Virginia Schools (West Virginia Board of Education)	2024
Generative AI Implementation Recommendations and Considerations for PK-13 Public Schools (North Carolina Department of Public Instruction)	2024
Human-Centered AI Guidance for K-12 Public Schools (Office of Superintendent of Public Instruction)	2024

Although the nine AI education guidelines provided general guidance in education regarding AI integration, our analysis identified a gap in the linking of SE industry AI usage to education. The gap between AI use in SE education and current industry practices with regards to AI provided the basis for the next steps in our research. The SE industry is using AI in ways that are not addressed in current curricula guidance.

2.2 Identification of Essential AI-Related SE Industry Segments

After identifying the lack of connection between SE industry AI usage and education, we conducted a systematic literature review research papers on AI use in the SE industry. The selection of these papers followed a pre-set criterion to ensure a comprehensive and unbiased review. We did a systematic online search using strings including "Research on Software Industry AI Use", "AI usage in the SE industry research", "AI usage in SE industry education" among others. From the initial search results, we screened papers based on the relevance to the SE industry, the presence of empirical data, and the discussion of AI applications (Table 2). Papers that did not meet these criteria were excluded.

Table 2: Criterion for selection process

Criterion	Description
Relevance to SE Industry	The paper must discuss AI applications within the context of the software engineering industry.
Empirical Data	The paper should include empirical data or case studies demonstrating the use of AI in SE.
Discussion of AI Applications	The paper should provide detailed information on AI technologies used in software development.
Evaluation of AI Tools	The paper must include methods and metrics for evaluating AI tools, such as code quality and productivity impacts.
Human-AI Interaction	The paper should examine interactions of human developers and AI tools.
Ethical and Privacy Considerations	The paper should discuss ethical implications and privacy concerns related to AI use in SE.
Impact on Skill Development	The paper should address how AI tools and concepts impact skill development and training in SE.

We determined that using guidance from the industry and its current AI usage based on research we could identify key segments impacted by AI on the SE industry.

Table 3: Research papers used in analysis

Title (Organization)	Year
Artificial Intelligence, Automation and Work (National Bureau of Economic Research)	2018
Systematic review of research on artificial intelligence applications in higher education – where are the educators? (International Journal of Educational Technology in Higher Education)	2019
Applications of AI in classical software engineering (AI Perspectives)	2020
Investigating Explainability of Generative AI for Code through Scenario-based Design (ACM)	2022
Productivity Assessment of Neural Code Completion (ACM)	2022
The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming (ACM)	2022
Developers talking about code quality (Empirical Software Engineering)	2023
Does Artificial Intelligence Promote or Inhibit On-the-Job Learning? Human Reactions to AI at Work (Systems)	2023
The Influence of Artificial Intelligence on Productivity in Software Development (Politecnico di Torino)	2023
From Copilot to Pilot: Towards AI Supported Software Development (University of Toronto & University of Victoria)	2023
Autonomous Intelligent Software Development (The Johns Hopkins Applied Physics Laboratory)	2023
The Rise of Intelligent System Development: A Qualitative Study of Developers' Views on AI in Software Development Processes (Umeå University)	2023

In the first round of analysis, we found 20 industry segments mentioned in some of the research papers we reviewed. Within each paper we used the frequency of each segment for determining essential segments (75% discussion threshold). To be deemed essential, segments must have been discussed in at least 75% of the research papers in our analysis. For example, Software Development was discussed in 12 of the 12 papers we reviewed while Development and Operations (DevOps) was discussed in 9 of the 12 papers (75% of all papers). We chose to use a 75% credibility threshold similar to the work done by Panagioti et al. (2019), where 75% indicate high heterogeneity, respectively. We were able to find if an industry segment was discussed in each paper by doing a search within each digital paper for each segment. Based on this criteria, these are the 10 essential SE industry segments defined:

Table 4: Top 10 Segments Deemed Essential

SE Industry Segment	Total Count	AI Integration/Use Description
Software Development	12	Automating code generation (& code completion), suggesting code improvements, or facilitating debugging processes.
Software Maintenance	12	Predicting software vulnerabilities, automating patches, or assisting in legacy code interpretation.
Process Improvement	12	Analyzing development processes, identifying bottlenecks, or suggesting optimizations.
Software Economics	12	Predicting costs, analyzing return on investment, or optimizing pricing strategies.
Knowledge Management	11	Organizing documentation, facilitating knowledge sharing, or automating information retrieval.
Project Management	10	Optimizing resource allocation, tracking project progress, or predicting project timelines.
Software Testing	10	Enhancing automated testing, generating test cases, or identifying potential points of failure.
Software Security	10	Detecting anomalies, predicting breaches, or automating security audits.

SE Industry Segment	Total Count	AI Integration/Use Description
Quality Assurance	10	Monitoring software quality, identifying regressions, or automating compliance checks.
Deployment and Operations (DevOps)	9	Optimizing deployment processes, managing infrastructure, or ensuring continuous integration/continuous deployment (CI/CD).

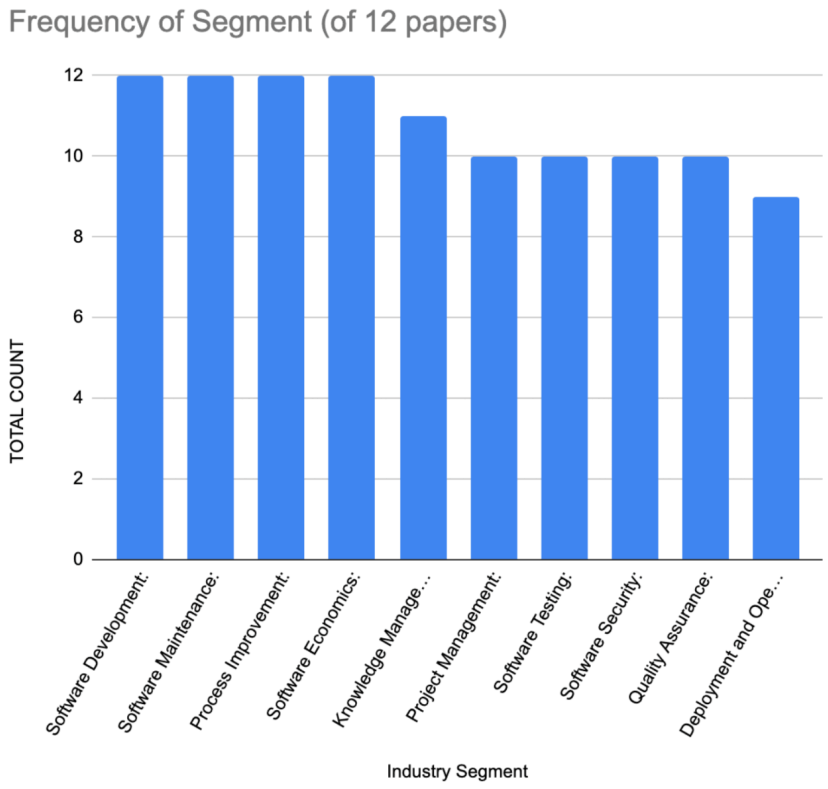


Figure 1: Frequency of top 10 segments in the 12 literature papers

By classifying each segment to each paper and storing the data in a spreadsheet, we were able to calculate the frequency of each segment overall. If there was no mention of a segment in a given paper, it is noted for reference. We were then able to calculate the frequency of each segment in the pool of our 12 papers in the literature review, listed in the graph above (Figure 1).

2.3 Identification of Essential AI-Related SE Industry SubSegments

In order to determine the subsegments, we conducted a similar process of identifying frequency of topics discussed in each of the segments by paper.

In our second round of analysis we based our framework on further insights from the research of the peer-reviewed papers in our analysis on AI-Usage in the SE Industry. We determined three sub-segments per segment to focus on in our framework based on frequency and expected impact on the industry through inductive content analysis. The total count refers to the number of times the topic appears in 1 of the 12 papers. For example, Code Generation, a sub-segment of Software Development, is mentioned in 10 of the 12 papers.

Table 5: AI-USE subsegments (AI use examples)

Top 10 Segments	AI-Use Sub-Segment 1	Total Count	AI-Use Sub-Segment 2	Total Count	AI-Use Sub-Segment 3	Total Count
Software Development	Code Generation	10	Code Completion	9	Software Prototyping	8
Software Maintenance	Automated Error Fixes	8	Code Refactoring	8	Bug Prediction	4
Process Improvement	Workflow Automation	12	Performance Optimization	12	Quality Control	9
Software Economics	Economic Impact Analysis	11	Resource Allocation	7	Cost Prediction	2
Knowledge Management	Semantic Reasoning	12	Automated Documentation	7	Knowledge Graphs	3
Project Management	Project Scheduling	10	Risk Assessment	8	Effort Estimation	5
Software Testing	Test Optimization	9	Fault Detection	5	Automated Test Generation	1
Software Security	Automated Patching	7	Security Risk Analysis	3	Vulnerability Detection	1
Quality Assurance	Code Quality Evaluation	10	User Feedback Analysis	10	Performance Benchmarking	6
Deployment & Operations (DevOps)	System Optimization	11	Monitoring & Alerts	5	Continuous Integration/Deployment	1

3. Expected Impact and Implications

We expect an enhancement of SE education through our AI-USE framework (Addison Lilholt, 2024). The AI-USE framework significantly enhances SE education by integrating AI concepts and applications into the curriculum. This framework is designed to cover critical areas of SE, including software development, maintenance, process improvement, economics, knowledge management, project management, testing, security, quality assurance, and DevOps. Each area is broken down into specific segments, providing detailed guidance on the use of AI tools and techniques.

For example, in software development, students learn about AI-driven code generation tools like GitHub Copilot and OpenAI Codex, which automate the creation of code snippets and templates, thus increasing efficiency and reducing manual coding efforts. Similarly, AI tools for code completion, such as IntelliCode and Kite, provide contextual suggestions that enhance coding speed and accuracy. In the segment of software maintenance, AI tools are employed for bug prediction, code refactoring, and automated error fixes. Predictive models trained on historical data can identify potential issues early, while refactoring tools and automated debugging systems ensure cleaner and more maintainable code. The framework also emphasizes AI's role in process improvement through workflow automation and quality control. By incorporating AI tools for cost estimation and resource allocation, the framework helps students understand how to predict and manage development costs and optimize resource use. Students learn that knowledge management is enhanced through the creation and use of knowledge graphs and automated documentation generation, improving knowledge sharing and decision-making. Project management is streamlined with AI-driven task management and risk assessment tools, which enhance task distribution and project security. In software testing, AI tools automate test case generation and execution, leading to faster and more comprehensive testing cycles. Security is bolstered by AI techniques for threat detection and vulnerability management, providing early detection of security threats and continuous system protection. Quality assurance benefits from AI-driven analysis of quality metrics and user feedback, ensuring higher software quality and user satisfaction. Finally, students are introduced to the segment of deployment and operations (DevOps) where AI tools enable faster release cycles and efficient resource utilization.

Overall, the AI-USE framework provides a comprehensive approach to enhancing SE education by equipping students with the knowledge and skills to leverage AI tools effectively, preparing them for the evolving demands of the software engineering industry.

The AI-USE framework addresses the evolving demands of the software engineering (SE) industry by integrating AI-driven tools and techniques into the educational curriculum, ensuring that students are well-prepared for the current and future outlook of the industry. It hopes to increase skill relevance and employability by familiarizing students with AI technologies, increasing their efficiency and productivity among other aspects of improvement.

The AI-USE framework hopes to prepare students for the evolving demands of the SE industry by equipping them with the latest AI tools and techniques, ensuring they can enhance productivity, improve quality, manage resources efficiently, while focusing on secure programming, software testing within CI/CD environments, and supply chain security, ensuring that students are equipped to address the critical security challenges posed by AI in software engineering. This comprehensive approach ensures that graduates are ready to meet the current and future challenges of the software engineering industry.

4. Goals for the Future

We hope to continue to develop our framework through various methods to ensure we provide high quality and applicable information moving forward. We hope to conduct Feedback Collection where we conduct interviews with SE Professionals from various segments of the SE industry to gather insights on the relevance and applicability of the AI-USE framework. The feedback will help us understand if our framework aligns with industry needs and where adjustments are necessary. This could be done through surveys and questionnaires to obtain quantitative data on the framework's effectiveness and areas for improvement.

We would also like to implement a standardized process for defining scope and impact where we would rank segments and subsegments as High, Medium, or Low impact based on feedback from industry professionals. This will help prioritize areas that need more focus and ensure the framework addresses critical aspects of the SE industry. We would also conduct an impact assessment to regularly assess the impact of each segment on the SE industry by analyzing feedback and industry trends. We could then adjust the framework to emphasize high-impact areas and address any gaps identified.

Finally, we would like to define the relationship to the Software Development Life Cycle (SDLC) to ensure that each segment and subsegment of the framework is mapped to specific phases of the SDLC. This will help educators and students understand the practical application of AI tools and techniques throughout the software development process.

5. Additional Methods for Continued Development

To further the development of the AI-USE Framework we would hope to conduct pilot programs and case studies where we implement pilot programs in collaboration with industry partners to test the AI-USE framework in real-world settings. According to research, pilot studies help identify potential issues, refine methodologies, and ensure that the framework can be effectively implemented on a larger scale. This process aids in the successful adaptation and scaling of educational initiatives (Institute of Education Sciences, 2023). We would document the outcomes and lessons learned to refine the framework. During this we would develop case studies based on successful implementations of AI tools in SE education. We would use these case studies as teaching materials to demonstrate the practical benefits of the framework.

It would be important to build academic-industry partnerships between academic institutions and industry to ensure continuous exchange of knowledge and best practices. These partnerships could provide students with internships and project opportunities that align with the AI-USE framework. Additionally the establishment of advisory boards composed of industry experts and academic leaders to provide ongoing guidance and support for the framework's development would be beneficial.

Regular updates and reviews should be made to the AI-USE framework to incorporate new AI technologies and methodologies. This will ensure that the framework remains current and addresses the latest developments in the SE industry. We would like to publish annual reports on the progress and impact of the framework, highlighting key achievements and areas for further improvement.

In addition we would like to assess feedback loops while conducting iterative improvement that allow for continuous input from users of the framework, including students, educators, and industry professionals using this feedback to make iterative improvements to the framework.

We have begun the development of an online platform where users can share their experiences and suggestions for enhancing the framework. This platform for ongoing development and collaboration. Such a platform can act as a community hub, allowing for collaboration and dissemination of best practices (Institute of Education Sciences, 2023).

By implementing these goals and methods, we anticipate the AI-USE framework remains a dynamic and valuable tool for SE education, effectively preparing students for the evolving demands of the software engineering industry.

6. Conclusion

The AI-USE framework addresses the global shortfall of software engineers by integrating AI concepts and applications into educational curriculum for middle and high schools. This strategic alignment aims to equip the future workforce with the necessary skills to meet the escalating demands of the SE industry, which is increasingly reliant on AI technologies. Our research involved a comprehensive review of nine international AI education guidelines and a systematic literature review of 12 research papers focusing on AI's role across the SE industry. Through this process, we identified 10 essential SE industry segments for inclusion in CS education: software development, software maintenance, process improvement, software economics, knowledge management, project management, software testing, software security, quality assurance, and deployment and operations (DevOps).

The AI-USE framework maps these essential segments to their predominant uses in the industry, providing detailed guidance on the application of AI tools and techniques and must also incorporate the necessary aspects of secure programming and software testing. This framework is designed to enhance SE education by covering critical areas, such as AI-driven code generation, bug prediction, workflow automation, cost estimation, and security threat detection. To ensure the framework remains relevant and effective, we propose an ongoing development process involving feedback from industry and academic professionals, regular updates, pilot programs, and academic-industry partnerships. By implementing these measures, the AI-USE framework will continuously evolve, preparing students to leverage AI effectively and meet the future challenges of the software engineering industry. This continued effort aims to create a future workforce that is more effective and capable of utilizing AI, potentially giving anyone using the framework a competitive edge globally.

References

Acemoglu, D. and Restrepo, P. (2018). Artificial Intelligence, Automation and Work. NBER Working Paper No. 24196. National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w24196

Barenkamp, M., Rebstadt, J. and Thomas, O. (2020). Applications of AI in classical software engineering. AI Perspectives, 2(1), 1-15. https://doi.org/10.1186/s42467-020-00005-4

Börstler, J., Bennin, K. E., Hooshangi, S., Jeuring, J., Keuning, H., Kleiner, C. et al. (2023). Developers talking about code quality. Empirical Software Engineering, 28(128). https://doi.org/10.1007/s10664-023-10381-0

Cardona, M. A., Rodríguez, R. J. and Ishmael, K. (2023). Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations. U.S. Department of Education. Retrieved from https://www2.ed.gov/documents/ai-report/ai-report.pdf

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B. and Barro, S. (2023). AI literacy in K-12: a systematic literature review. International Journal of STEM Education, 10(29). https://doi.org/10.1186/s40594-023-00418-7

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A. and Prather, J. (2022). The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming. In Australasian Computing Education Conference (ACE '22), February 14–18, 2022, Virtual Event, Australia. ACM. https://doi.org/10.1145/3511861.3511863

Garousi, V., Giray, G., Tüzün, E., Catal, C. and Felderer, M. (2019). Aligning software engineering education with industrial needs: A meta-analysis. Journal of Systems and Software, 157, 110393. https://doi.org/10.1016/j.jss.2019.06.044.

Glushkova, D. (2023). The Influence of Artificial Intelligence on Productivity in Software Development (Master's thesis). Politecnico di Torino. Retrieved from https://www.polito.it/handle/123456789/tesi/0001

Green, A., Elsey, T. and Crepaldi, N. (2023). The Great Acceleration: CIO Perspectives on Generative AI. MIT Technology Review Insights, sponsored by Databricks. Retrieved from https://www.databricks.com/blog/2023/04/13/retail-age-generative-ai.html

Holmes, W., Persson, J., Chounta, I.-A., Wasson, B. and Dimitrova, V. (2022). Artificial Intelligence and Education: A Critical View through the Lens of Human Rights, Democracy, and the Rule of Law. Council of Europe. Retrieved from https://rm.coe.int/prems-092922-artificial-intelligence-and-education-report-2022-eng/1680a7d36d

Li, C., Zhang, Y., Niu, X., Chen, F. and Zhou, H. (2023). Does Artificial Intelligence Promote or Inhibit On-the-Job Learning? Human Reactions to AI at Work. Systems, 11(3), 114. https://doi.org/10.3390/systems11030114

Paper Author. (2024). Educator Plans for AI-USE Framework. Retrieved from https://alilholt1.wixsite.com/ai-use-framework/educator-plans

Matties, M. A. (2023). Autonomous Intelligent Software Development. The Johns Hopkins Applied Physics Laboratory. Retrieved from https://arxiv.org/abs/2208.06393

Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W. and Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. Education Tech Research Dev, 71, 137–161. https://doi.org/10.1007/s11423-023-10203-6

North Carolina Department of Public Instruction. (2024). Generative AI Implementation Recommendations and Considerations for PK-13 Public Schools. Retrieved from https://go.ncdpi.gov/AI_Guidelines

Office of Superintendent of Public Instruction. (2024). Human-Centered AI Guidance for K-12 Public Schools. Retrieved from https://ospi.k12.wa.us/ai

Panagioti, M., Khan, K., Keers, R.N., Abuzour, A., Phipps, D., Kontopantelis, E., Bower, P., Campbell, S., Haneef, R., Avery, A.J. and Ashcroft, D.M., 2019. Prevalence, severity, and nature of preventable patient harm across medical care settings: systematic review and meta-analysis. BMJ, 366, p.l4185. Available at: http://dx.doi.org/10.1136/bmj.l4185

Pudari, R. and Ernst, N. A. (2023). From Copilot to Pilot: Towards AI Supported Software Development. University of Toronto & University of Victoria. arXiv:2303.04142

Institute of Education Sciences, 2023. *Pilot Study Resource*. [online] Available at: https://ies.ed.gov/ncee/edlabs/regions/appalachia/resources/pdfs/Pilot-Study-Resource_acc.pdf

Stavridis, A. and Drugge, A. (2023). The Rise of Intelligent System Development: A Qualitative Study of Developers' Views on AI in Software Development Processes. Umeå University. Retrieved from https://umu.diva-portal.org/smash/get/diva2:1759153/FULLTEXT01.pdf

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K. and Weisz, J. D. (2022). Investigating Explainability of Generative AI for Code through Scenario-based Design. In 27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland. ACM. https://doi.org/10.1145/3490099.3511119

UNESCO. (2021). AI and Education: Guidance for Policy-makers. Miao, F., Holmes, W., Huang, R. and Zhang, H. https://doi.org/10.54675/PCSP7350

West Virginia Board of Education. (2024). Guidance, Considerations, & Intentions for the Use of Artificial Intelligence in West Virginia Schools. Retrieved from https://wvde.us/ai-guidance

Yim, I. H. Y. and Su, J. (2023). Artificial intelligence (AI) learning tools in K-12 education: A scoping review. Journal of Computers in Education. https://doi.org/10.1007/s40692-023-00304-9

Zawacki-Richter, O., Marín, V. I., Bond, M. and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Ziegler, A., Kalliamvakou, E., Li, X. A., Rice, A., Rifkin, D., Simister, S., Sittampalam, G. and Aftandilian, E. (2022). Productivity Assessment of Neural Code Completion. In Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming (MAPS '22), June 13, 2022, San Diego, CA, USA. ACM. https://doi.org/10.1145/3520312.3534864

Empowering Diversity by Building Inclusive Software Engineering Projects with Large Language Models

Clara Maathuis[1], Greg Alpar[2] and Stefano Bromuri[1]
[1]Open University of the Netherlands, Heerlen, The Netherlands
[2]Radboud University, Nijmegen, The Netherlands

clara.maathuis@ou.nl
greg.alpar@ru.nl
stefano.bromuri@ou.nl

Abstract: The integration of emerging AI technologies, particularly Large Language Models (LLMs), is fundamentally reshaping the landscape of software engineering. LLMs offer a wide array of capabilities that can enhance various aspects of software development, including design assistance, automated code analysis and synthesis, and testing. Consequently, their integration into software engineering practices holds the potential to greatly improve efficiency and code quality, marking a notable paradigm shift towards the creation of more intelligent and adaptive systems. This allows the definition of an extended perspective when establishing requirements for building software engineering solutions by incorporating evidence from literature resources and practice. By synthesizing requirements and evidence from scholarly and practitioner sources using LLMs, software engineers can ensure that their solutions are not only technically sound but also align with best practices in fostering social norms and values like inclusivity. Adopting such an approach supports the creation of responsible and inclusive educational software that caters to diverse learning needs and promotes equitable access to educational resources. Moreover, by using LLMs to inform decision-making throughout the software development lifecycle, software engineers can iteratively refine and enhance their solutions based on emerging research findings, thereby ensuring continuous improvement in fostering inclusive educational environments. Hence, this research aims to develop a novel evidence-based software engineering method informed by insights from scientific literature. As a use case, we design and implement a dyslexia-oriented educational software application that supports children in learning to read, guided by this new methodological approach.

Keywords: Evidence-based software engineering, AI, LLMs, Diversity, Neurodiversity, Dyslexia

1. Introduction

"Make reading a habit, and it will become a source of pleasure and knowledge." (Isaac Asimov)

The integration of LLMs in software engineering is transforming the software development process by enhancing various activities, including design assistance, automated code synthesis, and testing. This process is fostering a paradigm shift towards the development of more intelligent and adaptive systems, necessitating the exploration of novel methods and techniques that redefine traditional software development practices. At the same time, it is essential to prioritize societal well-being and balance ethical values like diversity, inclusivity, and social responsibility to ensure that software systems are beneficial, equitable, and sustainable. In educational settings, addressing diversity in software engineering projects through LLMs requires a nuanced and multidisciplinary approach to understand and tackle the multifaceted nature of this context. By focusing on developing inclusive software engineering processes with LLMs, it is possible to capture and implement the essential elements necessary for effective software development, leading to the creation of software systems that are not only efficient, but also socially responsible and sustainable (Huang et al., 2024; Bach et al,, 2022; Amatriain, 2024; Shah, 2024).

For children and individuals with dyslexia, the struggle with reading is often a daily and frustrating experience. When faced with a written text, their brains have difficulty processing the visual information, making it challenging to decode letters, to recognize words, and comprehend the meaning of the text. This can lead to feelings of anxiety, embarrassment, and low self-esteem. Children with dyslexia may experience difficulty sounding out words, remembering letter sequences, and understanding sentence structure, making reading a laborious and exhausting task (Mather, White & Youman, 2020; Akyurek et al., 2024). Further, impact can be observed on their academic performance, social interactions, and overall confidence, making it essential to provide targeted support to help them overcome their reading challenges. While various systems consider the integration of dyslexia-friendly supportive methods or are directly dedicated to dyslectic users (Akyurek et al., 2024), still, a limited number of studies and existing solutions are built using advanced AI technologies. To address this, we build a novel dyslexia-oriented software application prototype using LLMs, prompt engineering, and RAG (Retrieval Augmented Generation) based on extensive multidisciplinary literature review in an evidence-based software engineering methodological approach. To the best of our knowledge, this

represents the first attempt in the literature in this direction and is based on an unique approach that can benefit from the development of state-of-the-art AI technologies applied in this domain. Building software engineering systems with Large Language Models (LLMs) in the context of dyslexia is important because it enables the creation of tailored, adaptive, and user-friendly tools that can assist children with dyslexia in overcoming their reading challenges, improve literacy skills, and increase their confidence. Hence, by leveraging LLMs, software engineers can design inclusive educational tools that support diverse learning needs, exemplifying the challenges and goals of inclusive software engineering.

The outline of the article is structured as follows. Section 2 discusses the technical dimension of the research background. Section 3 presents related studies that address diversity in the software engineering domain using advanced AI technologies such as LLMs. Section 4 presents the methodological approach adopted in this research. Section 5 introduces the design and development of the prototype proposed. At the end, concluding remarks and future research perspectives are provided.

2. Research Background

Software engineering encompasses the process of developing software to fulfil specific customer requirements. Initially, the customer or software developer identifies the problem and outlines the desired characteristics and features. Over time, changes to these requirements may be suggested and implemented. With a history that can be traced back to the 1960s when the advent of integrated circuits revolutionised the computer market, software engineering involves a broad range of activities that collectively constitute the software development process. These activities encompass requirements gathering, system design, software development, testing, deployment, maintenance, documentation, project management, collaboration and communication, and quality assurance (Royce, 1970; Boehm, 1979). Particularly in Agile development methodologies, these steps are often iterative and may overlap (Abrahamsson et al., 2017). The ultimate objective is to deliver a functional and reliable software system that effectively caters to the needs of its users. The advancement of software technology played a critical role in driving the progress of software engineering. Numerous methods and techniques were developed to facilitate efficient and dependable software production, and various programming environments provide integrated tools like compilers, debuggers, and syntax-oriented editors to support programming tasks. In the realm of modern software engineering, the continuous progress of software technology continues to shape the field, introducing novel methods and technologies that enhance the efficiency and reliability of software production. This includes practices like Agile development methodologies (Abrahamsson et al., 2017), DevOps, cloud computing, and AI, which revolutionised the entire software development, deployment, and maintenance processes (Bourque & Fairley, 2014).

An emerging paradigm that aims to integrate research findings and empirical evidence into software development practices aiming to support various stakeholder needs is named evidence-based software engineering (EBSE). This involves systematically collecting and analysing empirical data to support decision-making processes throughout the software development lifecycle, and further allow software engineers to make informed choices, mitigate risks, and improve the overall quality and efficiency of software projects (Kitchenham, Dyba & Jorgensen, 2004). This approach draws inspiration from the evidence-based practices established in other disciplines like psychology, medicine, and manufacturing while recognising the importance of basing software engineering decisions on empirical evidence rather than relying solely on expert knowledge and experience (Kitchenham, Budgen & Brereton, 2015). This facilitates the translation of research findings into practical design requirements and guidelines that can be applied in real-world contexts (Huang et al., 2024). Given the recent developments in the AI domain that include LLMs, a new avenue is opened for EBSE systems given the intrinsic ability of LLMs to process and generate human-like text that can be applied for various EBSE tasks like literature review, data extraction, and synthesis of requirements and guidelines formulation based on various data sources. LLMs can be fine-tuned on domain-specific datasets using prompt engineering techniques (Hadi et al., 2023; Minaee et al., 2024) to improve their performance in identifying and extracting key findings and empirical results from sources such as research articles and field standards and documentation. Prompt engineering involves providing the input prompts that guide the LLM towards the desired task, making use of its understanding capabilities (Bach et al,, 2022; Amatriain, 2024; Shah, 2024). Furthermore, RAG (Retrieval Augmented Generation) can be employed to augment the LLM's knowledge with external information retrieval from relevant field sources and provide more accurate and comprehensive evidence synthesis (Lewis et al., 2020; Feng et al., 2024). Hence, the integration of prompt engineering and RAG techniques into EBSE development has the potential to enhance the quality and reliability of the evidence extraction and synthesis process. In this way features and patterns are identified and translated to design

requirements that are further accounted for implementation in the EBSE process. This mechanism can be applied following two strategies. First, extracting and summarizing the evidence that is provided to the LLM as a short set of requirements. And second, by downloading relevant evidence sources (e.g., research articles, technical and domain documentation) and implementing RAG to provide requirements through prompt engineering. In this research, both strategies are applied.

3. Related Studies

To sustain the ongoing progress of society in a fair, just, and sustainable manner, the diversity, equity, and inclusivity principles and values need to be acknowledged, respected, implemented, and promoted at local and global scale as well as at individual and collective levels, when necessary, through dedicated and proactive strategies (Gill et al., 2018). While diversity implies the recognition, appreciation, and celebration of the whole spectrum of human differences existing in society, equity relates to providing equal opportunities, resources, and access in the same way to all individuals. This is especially important for those who have historically experienced marginalisation. Inclusion refers to building and maintaining an environment where all individuals are respected, valued, and included, regardless of their particular characteristics or backgrounds (UN, 2019; ISC2, 2022; Russen and Dawson, 2023). These principles are strongly connected and contribute to addressing global challenges and to achieving sustainable goals related to the quality of education, gender equality, promotion of decent work and economic growth, reducing inequalities, and promoting peace, justice, and strong institutions (UN, 2015; Schwarz, 2023). Particularly, diversity, equity, and inclusion are essential in both education and work contexts for promoting the achievement of educational learning goals and work-related objectives, student or individual well-being while building and maintaining a positive social and educational development (Claeys-Kulik et al., 2023; Barnett, 2020).

In an educational context, assuring a diverse, equitable, and inclusive learning environment is crucial. Such an environment fosters exposure to diverse perspectives and experiences, enriching students 'learning journey. Promoting interaction among students and with teachers cultivates empathy and respect within the educational community. Moreover, possible human and educational disparities that could occur based on aspects related to race, social status, or disabilities, are reduced while building and sustaining a positive and welcoming educational environment for students and staff through a general sense of community, acceptance, value, and respect that can only enhance the overall learning environment (UNESCO, 2017; Gretter et al., 2019; Washington, 2020). Concurrently, as part of the learning environment are the systems and tools used for supporting learning activities, processes, and overall communication. Designing educational systems with consideration for diversity, equity, and inclusion is essential for creating effective, efficient, and responsible educational programs that cater to the diverse needs and backgrounds of learners. (Fraser and Mancl, 2023). This underscores the value of inclusivity in promoting learning outcomes and preparing students for a diverse world Furthermore, embedding and respecting these principles also results to software that is relevant and usable to a wider range of users. This implies an improved user experience, more informed and effective decision-making benefiting from a broader range of viewpoints and insights, and an increased user engagement, creativity, and innovation (Jurado de Los Santos et al., 2020; Albusays et al., 2021; UN, 2022a; UN, 2022b).'

In the area of computer programming, Biswas (2023) investigates the role and capabilities of ChatGPT to provide users assistance, explanations, examples, and guidance for understanding programming concepts and techniques, finding useful resources, and solve specific technical problems. Accordingly, this study crosses a broad range of examples in Python and illustrates how ChatGPT could assist in solving problems in an effective way. In these lines, Biswas (2023) analyses the use of ChatGPT in Data Science projects to automate various project aspects such as data cleaning and pre-processing, model training, and interpretation of results. In higher education and professional development domains, Hassani and Silva (2023) conducts a study making use of GPT-3 to generate content on the study material and to provide a series of examples and prompts for tasks including automated essay scoring, research assistance, language translation, syllabus creation, quizzes, and exams for teaching purposes, and to generate summaries, reports, and other research-related instruments. In the context of using ChatGPT for learning and assessment of computer science curriculum at undergraduate level, Qureshi (2023) conducts an experimental study with two groups of students with a total number of 20 students whose performance was measured with and without making use of ChatGPT. The students participating in this study had to solve diverse tasks related to data structures and algorithms; for instance, sorting arrays, graph topological sorting, and Dijkstra's algorithm. The results suggest that while ChatGPT is able to produce coherent and well-structured code, it generates promising outputs and

comprehensive explanations. However, there is a need for a step-by-step approach to being guided by it in order for students to benefit from its guidance and to receive useful and well-explained results.

Wang et al., (2024) review LLMs comprehensively in the education domain to address various applications and benefits compared to traditional methods. The study categorizes educational applications based on user roles and scenarios, highlighting the transformative role of LLMs in education. The survey outlines datasets and benchmarks used for evaluating LLMs in educational settings, including question construction and automatic grading tasks. The authors further emphasize the importance of diversity in educational settings and stresses the challenges faced by certain demographic groups due to unequal access to educational models of comparable quality levels, highlighting the need for inclusivity. This directly relates to issues like social and language biases, fairness, and transparency of the models used (Maathuis & Chockalingam, 2024). These models need to be designed inclusively to accurately reflect the diversity of the student population and assure equitable learning opportunities for all students. In line with this, Dongre et al., (2024) explore ways to enhance empathy in LLMs by integrating psychological data and building the Empathic LLM (EmLLM) chatbot for stress monitoring and control in order to provide human-like responses and assessment. For neurodivergent individuals, Choi et al., (2024) emphasize the benefits of integrating conversational agents into their lives as they would be able to provide tailored responses to their needs, further stressing the risk of over-reliance on such solutions. From here, the authors recommend fostering autonomy and agency while respecting the unique needs of users. In a practical setting, Jang et al., (2024) simulate interactions between autistic individuals and LLMs revealing that many participants show a strong preference for LLM-generated over human advice, highlighting the potential for increased agency in navigating social situations. Nevertheless, the authors stress their concerns regarding the LLMs' tendency to predict best-case scenarios, potentially leaving individuals unprepared for unexpected interactions, and they highlight the importance of adopting an assistive framing in developing such technologies that truly serve the needs and autonomy of autistic individuals.

In the context of dyslexia, Akyurek et al., (2024) conduct a systematic review on current methods that improve cognitive skills such as reading skills of children, and recall promising methods like Sports-Vision Exercises, attention training, phonology reading writing programs, auditory and visual timing interventions, reading with masking interventions, working memory training, virtual reality neurorehabilitation treatments, and software programs like MAEVA(c) and RapDys(c). These methods target various aspects of reading including word accuracy, comprehension, and phonemic perception. Dyslexia is primarly caused by an imbalance in the functioning of the left brain hemisphere, and its core features can be classified as follows (Shaywitz, 1996; Egan & Tainturier, 2011; Savill & Thierry, 2011; Roitsch & Watson, 2019; Mather, White & Youman, 2020; Ahire et al., 2023):

- Reading and writing challenges: difficulty involving challenges in reading, word-spelling, writing, reasoning, word decoding, and other neurological traits regardless of intelligence level expressed through slow reading and writing skills, confusion among words, difficulty in reading single words in isolation without context, spelling errors, struggles with tasks, and incorrect perception of alphabets;
- Cognitive and processing difficulties: complications in various areas such as phonological, orthographic, working memory, brain systems asynchrony, poor executive function skills, and rapid naming processing;
- Language specific variations: variations per language as the linguistic demands vary and in languages like Chinese and Arabic, memory, orthographic knowledge, phonological awareness, and rapid naming play crucial roles in reading accuracy and fluency;
- Development and age-related differences: variations presents as individuals age, with young children struggling with sound-letter associations and older individuals facing issues with reading fluency, spelling, and written expression, which persist into adulthood.

As our review shows, various software engineering and LLMs-based solutions have been proposed for tackling various diversity aspects. Regarding dyslexia, to the best of our knowledge, while relevant solutions have been proposed for providing educational support in activities such as reading while reflecting on the potential impact produced, there exists a knowledge gap related to the incorporation of LLMs-based approaches for building an assistive intelligent solution for children with dyslexia.

4. Research Methodology

This research aims to build a software application with LLMs for reading assistance for children with dyslexia. On this behalf, the following research questions (RQ) are formulated:

RQ1: *What are the main features that characterize the reading process of children with dyslexia?*

RQ2: *What are the design requirements considered in the Prompt Engineering process?*

RQ3: *How can a software application be built by means of LLMs for reading assistance for children with dyslexia?*

To achieve this objective, an evidence-based software engineering approach is considered and implemented (Kitchenham, Budgen & Brereton, 2015; Huang et al., 2024) as depicted in Figure 1 below.

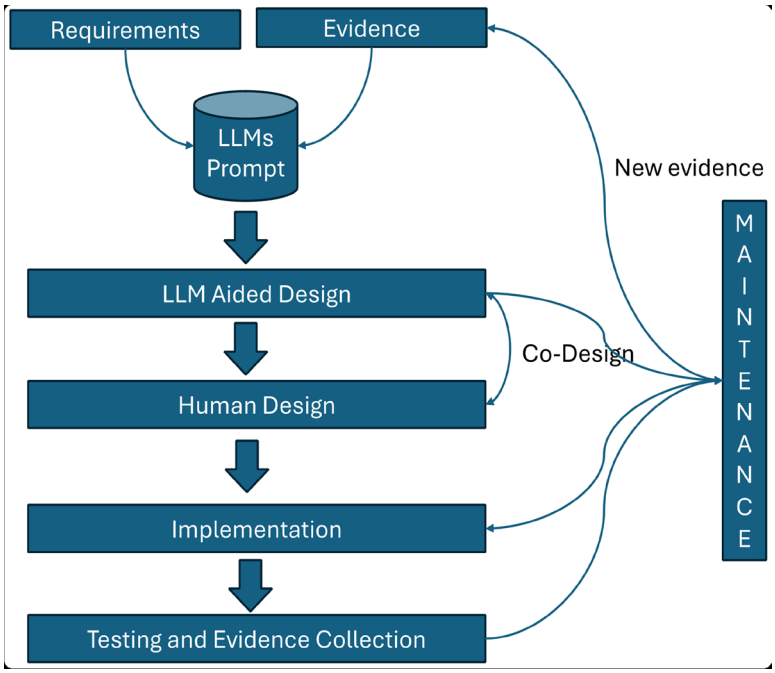


Figure 1: Methodological approach

In this process, the evidence is gathered from scientific sources and prompts are created to guide the LLM in the initial phase. Further, a collaborative co-design takes place, in which the human developers and the LLM iteratively refine and integrate responses from the LLM to develop the application in a unique human-AI teaming setting. At the end, the human developers take the lead by building the application integrating LLM-generated components and conducting initial tests to assess its functionality. After deployment, further refinement and improvements are possible based on the operation and usage of the application. In the next section, these phases are elaborated.

5. Solution

In **Phase 1 (Literature review)**, given the multidisciplinary nature of this research, an extensive literature review was conducted in the AI, software engineering, education, medicine, psychology, and diversity domains. In this process, combinations of keywords like *diversity, dyslexia, evidence-based software engineering, prompt engineering, LLMs, and ethics* were used to query scientific databases like IEEE Xplore, ACM Digital Library, Springer, Wiley, and Google Scholar.

In **Phase 2 (Experimentation with prompt engineering),** the collected studies were analysed and the translation of research findings into design requirements and implementation is executed through the following mechanisms:

- **Zero-shot prompting** where instructions are provided to ChatGPT 3.5 without examples or demonstration. In this way, the model generates its response based on its training data and task understanding. The prompt provided is as follows:

act as a software engineer and AI expert, write a software program in Python that is teaching reading.

The results provide a general program together with corresponding explanations that point to the fact that further functionalities like user progress tracking or a word database could be added.

- **Few-shot learning** where a small number of examples are provided to the model as part of the prompt to guide the model towards generating the desired output. In this case, the dyslexia context was given to the model using the prompt:

act as a software engineer and AI expert, write a software program in Python that is teaching reading taking into consideration dyslexia.

The results provide a program that mentions the incorporation of dyslexia-friendly font or style for displaying words.

- **Chain-of-thought prompting** where examples that demonstrate the step-by-step reasoning process leading to the solution are provided in the prompt. Here, features gathered synthetizing the findings obtained from the extensive literature review were provided to the model. The features contain textual, screen, sound, and educational elements, as follows:

act as a software engineer and AI expert, write a software program in Python that is teaching reading taking into consideration dyslexia. For this consider as features of the program the following ones: couple reading with audio, reading is preceded by phonological training, offer breaks to the users, consider adjustable timing, or even omit timing when reading, consider that the context of a word plays an important role, and distinguish between diagnosis and intervention.

The results provide a program that takes into consideration a part of these features and further reflecting on each of the category of features considered while pointing out on the possibility of including educational experts, speech therapists, and psychologists in such as project.

- **RAG** where the model is requested to retrieve relevant information from the research knowledge sources provided complementary to its internal knowledge. This retrieved information is then combined with the prompts provided by human developers to generate more informed and contextually relevant responses with respect to the design and development of the solution. Further, examples of prompts used are provided:

describe the relationship between dyslexia and reading.

describe tests to identify dyslexia.

describe effective approaches to train someone with dyslexia to read.

describe an application to learn how to read with dyslexia.

The results provide valuable information about the features considered and a series of reflective point for each category of features. Nevertheless, based on the human assessment conducted, while no significant points were found, the results are proportional to the level of evidence provided. This means that with a more multidisciplinary or transdisciplinary systematically analyzed evidence, valuable differences could potentially be seen.

In **Phase 3 (Solution development),** a prototype web application is built to assist children with dyslexia to read based on the evidence found and the LLM support provided. It is important to note that in this process, the use of RAG did not provide additional findings when compared to the shot-based approach. Accordingly, a preview of the website is provided in Figure 2.

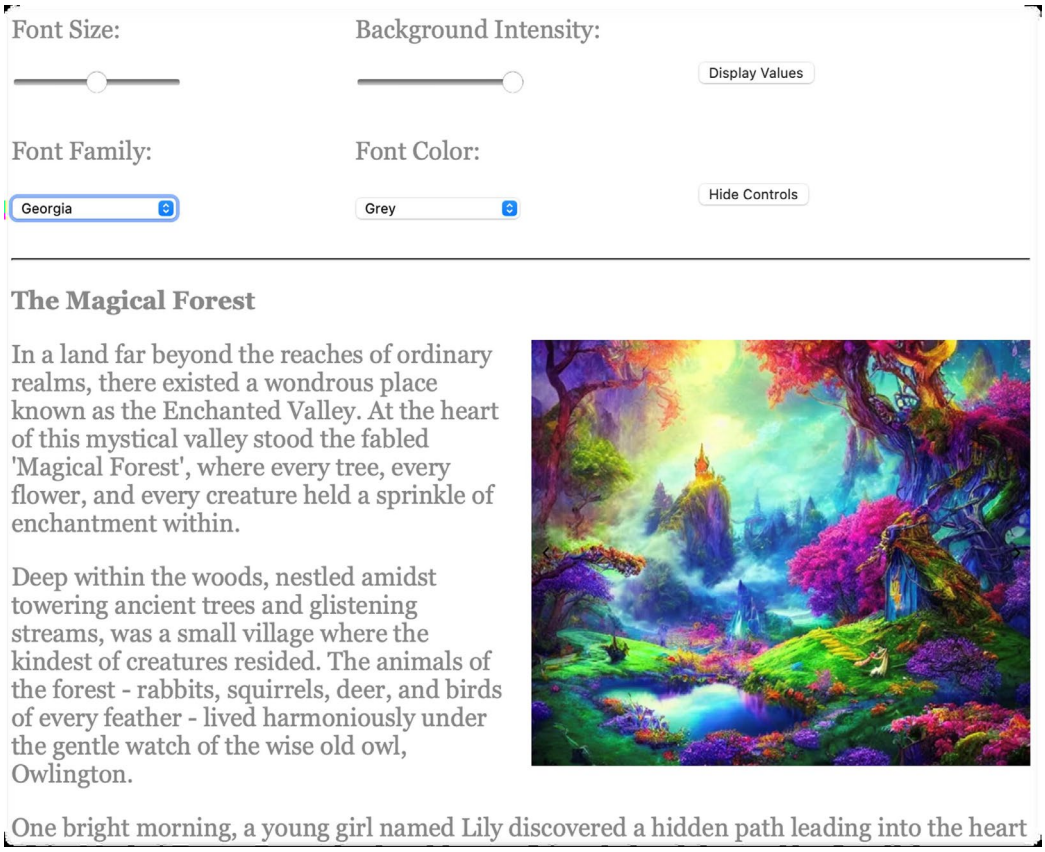


Figure 2: Prototype web application

The web application presents a fictional story and integrates a series of font size, type, and colour as well as background intensity controls that allow users to experiment and assess which combination is helpful in their reading process and assure an inclusive reading experience. As this represents a prototype, the intention is to evaluate it with experts and children. Based on the feedback received, the solution will be further adapted accordingly in future research.

6. Conclusions

This research introduces a novel evidence-based software engineering methodology designed to address the complexities of developing inclusive educational tools. By applying this new method, we aim to create innovative software solutions that cater to diverse learning needs, particularly for children (and adults) with dyslexia who have historically faced challenges in accessing educational resources. The methodology integrates scientific insights and practical knowledge while maintaining a user-centric perspective that prioritizes the specific needs of this audience. Our approach includes identifying the key features that characterize the reading process for children with dyslexia, establishing design requirements for effective prompt engineering, and ultimately developing a software application that provides tailored reading assistance to enhance their learning experience.

As future research perspectives, this research continues by conducting the evaluation of the proposed prototype with experts in dyslexia and children with dyslexia for refining and improving purposes. This evaluation process would not only ensure the application's effectiveness and usability, but would also provide valuable insights for future enhancements and research directions. Incorporating feedback from the target users is essential in creating an inclusive user-centric and impactful solution. Further developments aim to evolve the solution into a comprehensive and adaptive platform that addresses the diverse needs of children and dyslexia across different languages, cultures, and learning preferences. The evaluation will ultimately foster more inclusive and equitable educational landscape.

References

Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2017). Agile software development methods: Review and analysis. *arXiv preprint arXiv:1709.08439*.

Ahire, N., Awale, R. N, Patnaik, S. & Wagh, A. (2022). A comprehensive review of machine learning approaches for dyslexia diagnosis. *Multimedia Tools and Applications*, 82.

Albusays, K., Bjorn, P., Dabbish, L., Ford, D., Murphy-Hill, E., Serebrenik, A., & Storey, M. A. (2021). The diversity crisis in software development. *IEEE Software*, *38*(2), 19-25.

Amatriain, X. (2024). Prompt design and engineering: Introduction and advanced methods. *arXiv preprint arXiv:2401.14423*.

Akyurek, G., Koca, R. B., & Gunal Gunser, R. (2024). Systematic review of current approaches to cognitive skills in children with dyslexia. *Current Psychology*, 1-17.

Bach, S. H., Sanh, V., Yong, Z. X., Webson, A., Raffel, C., Nayak, N. V., ... & Rush, A. M. (2022). Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*.

Barnett, R. (2020). Leading with meaning: Why diversity, equity, and inclusion matters in US higher education. *Perspectives in Education*, *38*(2), 20-35.

Biswas, S. (2023). Role of ChatGPT in Computer Programming.: ChatGPT in Computer Programming. *Mesopotamian Journal of Computer Science*, *2023*, 8-16.

Boehm, B. W. (1979). Software Engineering-as it is. IEEE ICSE 1979, pp. 11-21.

Bourque, P. & Fairley, R. E. (2014). Guide to Software Engineering Body of Knowledge version 3.0. IEEE Computer Society.

Claeys-Kulik, A. L., Jørgensen, T. E., & Stöber, H. (2019). Diversity, equity and inclusion in European higher education institutions. *Results from the INVITED Project. Brussel: European University Association Asil, 51p*.

Choi, D., Lee, S., Kim, S. I., Lee, K., Yoo, H. J., Lee, S., & Hong, H. (2024, May). Unlock Life with a Chat (GPT): Integrating Conversational AI with Large Language Models into Everyday Lives of Autistic Individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-17).

Dongre, P., Behravan, M., Gupta, K., Billinghurst, M., & Gračanin, D. (2024). Integrating Physiological Data with Large Language Models for Empathic Human-AI Interaction. *arXiv preprint arXiv:2404.15351*.

Egan J. & Tainturier M. J. (2011) Inflectional spelling deficits in developmental dyslexia. *Cortex*, 47(10):1179-96

Feng, Z., Feng, X., Zhao, D., Yang, M., & Qin, B. (2024). Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 11661-11665). IEEE.

Fraser, S., & Mancl, D. (2023). Dimensions of Diversity, Equity, and Inclusion. *ACM SIGSOFT Software Engineering Notes*, *48*(2), 18-21.

Gill, G. K., McNally, M. J., & Berman, V. (2018, September). Effective diversity, equity, and inclusion practices. In *Healthcare management forum* (Vol. 31, No. 5, pp. 196-199). Sage CA: Los Angeles, CA: SAGE Publications.

Gretter, S., Yadav, A., Sands, P., & Hambrusch, S. (2019). Equitable learning environments in K-12 computing: Teachers' views on barriers to diversity. *ACM Transactions on Computing Education (TOCE)*, *19*(3), 1-16.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. *Big data and cognitive computing*, *7*(2), 62.

Huang, Y., Chen, Y., Chen, X., Chen, J., Peng, R., Tang, Z., ... & Zheng, Z. (2024). Generative Software Engineering. *arXiv preprint arXiv:2403.02583*.

ISC2, 2022. 80 diversity and inclusion definitions. ISC2 Toolkit .

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004). Evidence-based software engineering. In *Proceedings. 26th International Conference on Software Engineering* (pp. 273-281). IEEE.

Jurado de Los Santos, P., Moreno-Guerrero, A. J., Marín-Marín, J. A., & Soler Costa, R. (2020). The term equity in education: A literature review with scientific mapping in web of science. *International Journal of Environmental Research and Public Health*, *17*(10), 3526.

Jang, J., Moharana, S., Carrington, P., & Begel, A. (2024, May). "It's the only thing I can trust": Envisioning Large Language Model Use by Autistic Workers for Communication Assistance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Kitchenham, B. A., Budgen, D., & Brereton, P. (2015). *Evidence-based software engineering and systematic reviews* (Vol. 4). CRC press.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Maathuis, C., & Chockalingam, S. (2024). Risk Assessment of Large Language Models Beyond Apocalyptic Visions. In *European Conference on Cyber Warfare and Security* (Vol. 23, No. 1, pp. 279-286).

Mather, N., White, J. & Youman, M. (2020). Dyslexia around the world: a snapshot

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Qureshi, B. (2023). Exploring the use of chatgpt as a tool for learning and assessment in undergraduate computer science curriculum: Opportunities and challenges. *arXiv preprint arXiv:2304.11214*.

Rotisch, J. & Watson, S. M. (2019) An overview of dyslexia: definition, characteristics, assessment, identification, and intervention. *Science Journal of Education*, &(4).

Royce, W. W. (1970). Managing the development of large software systems.

Russen, M., & Dawson, M. (2023). Which should come first? Examining diversity, equity and inclusion. *International Journal of Contemporary Hospitality Management*, *36*(1), 25-40.

Savill, N. J., & Thierry, G. (2011). Reading for sound with dyslexia: Evidence for early orthographic and late phonological integration deficits. *Brain Research, 1385,* 192–205

Shaywitz S. E. (1996). Dyslexia. *Sci Am.,* 275(5):98-104.

Schwarz, S. K. (2023). Equity, Diversity, and Inclusion in the Journal and the United Nations Sustainable Development Goals. *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, *70*(6), 925-929.

Shah, C. (2024). From Prompt Engineering to Prompt Science With Human in the Loop. *arXiv preprint arXiv:2401.04122*.

UN (2015). What are the sustainable development goals. UN URL: https://www.undp.org/sustainable-development-goals.

UN (2019). United nations disability inclusion strategy. UN .

UN (2022a). European declaration on digital rights and principles for the digital decade. UN .

UN (2022b). Towards equity and inclusion in higher education in europe. UN .

UNESCO (2017). A guide for ensuring inclusion and equity in education. UNESCO.

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., ... & Wen, Q. (2024). Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Washington, A. N. (2020). When twice as good isn't enough: The case for cultural competence in computing. In *Proceedings of the 51st ACM technical symposium on computer science education* (pp. 213-219).

Navigating Online Narratives on Israel-Hamas War with LLMs

Clara Maathuis and Iddo Kerkhof
Open University of the Netherlands, Heerlen, The Netherlands

clara.maathuis@ou.nl
email@iddo.eu

Abstract: In recent years, international conflicts have seen a steady rise, accompanied by the increased use of technological advancements to achieve strategic and military objectives. One such conflict, the ongoing hostilities between Israel and Hamas, exemplifies the complex interplay of political, ideological, social, and technological factors shaping contemporary warfare. While extensive analysis is devoted to analysing strategic, historical, and economic dimensions, still limited attention is directed through research efforts towards the vast troves of information available on global social media platforms such as YouTube. These platforms serve as conduits for a wide range of narratives, opinions, and representations related to this war, by providing valuable insights into public discourse and sentiments. Given the capability of automatically and systematically analysing vast amount of content, this research tackles the above stated knowledge gap by developing a novel explorative LLMs (Large Language Models)-based modelling approach to efficiently extract, categorize, and interpret nuanced discussions surrounding the Israel-Palestinian war. It does that following the Data Science methodological approach and focusing on YouTube data aiming at not only contributing to enriching understanding of contemporary conflicts through social media platforms, but also to reflect on the evolving role that state-of-the-art AI technological developments play in shaping the global security landscape.

Keywords: Israel-Hamas war, Israel-Palestine War, Generative AI, Large language models, LLMs, Knowledge graphs

1. Introduction

> *"Artificial intelligence is a tool that can be used for good or bad. It is not good or bad in itself. It is how we use it." (Andrew Ng)*

The ongoing Israeli-Palestinian conflict, with its deep historical roots, continues to have profound and far-reaching consequences. The multifaceted impacts of this war imply both societal and strategic struggle as well as widespread chaos and hardship (Buheji & Hasan, 2024). At the same time, people deal further with a wide spectrum of psychological and mental health effects including post-traumatic stress disorder (PTSD), anxiety, and depression, directly linked to prolonged exposure to violence, destruction, and loss (Abudayya et al., 2023). In this context, building intervention strategies implies understanding the meaning and implications of such actions as well as how solutions for resilience, conflict resolution, and peace-building mechanisms could be thought and actually put in place. At the same time, diverse demographics shape perspectives across different countries, with age influencing empathy levels and contributing to the growing global empathy for Palestine amidst ongoing political and social challenges (Buheji, Mushimiyimana & Ahmed, 2024). Understanding the nature of this war is crucial for applying the correct legal principles and comprehending the humanitarian impact on the civilian population involved. In this sense, the multifaceted efforts of the global strategic audience to assist through governance, legal, and social instruments, reveal the need for a comprehensive, integrated approach that addresses the diplomatic, legal, humanitarian, and psychological dimensions of this war (Akgül-Açıkmeşe & Özel,, 2024; ICC, 2024).

In this context, through social media platforms, the global narratives and public opinion are (re)shaped as these contribute to cultural communication and opinion formation within this complex geopolitical context. These digital spaces have emerged as critical arenas for information dissemination, digital activism, and the moulding of public perception (Anisa, Harjoko & Khadifa (2023). Then a nuanced understanding on the narratives embedded in social media is necessary as it plays a significant role in information propagation, public sentiment formation or crystallization, and mobilization support. This digital dimension introduces an additional layer of complexity to the conflict, where information warfare and narrative control operate concurrently with physical hostilities (Dostri, 2024). In this arena, AI (Artificial Intelligence) demonstrated important success in both building analytical, content generation, and user experience solutions as well as building mechanisms of social media manipulation and efforts of countering it (Fard & Maathuis, 2024), LLMs started to emerge as a powerful force in this domain. Taking into account the fact that LLMs have shown remarkable versatility across various applications, including language translation, text generation, and question answering. In the context of social media, these capabilities translate into enhanced content creation, more sophisticated chatbots, improved sentiment analysis, and more accurate content moderation. The social impact of LLMs in social media is multifaceted, presenting both challenges and opportunities (Ya et al., 2023: Maathuis & Chockalingam, 2024). As noted by Myers (2024), public scepticism surrounding AI implementation

in social media, particularly concerning privacy and data security, necessitates careful management and transparent communication. The potential for LLMs to perpetuate or amplify biases present in their training data is a critical concern, especially in the context of social media where such biases could have far-reaching societal impacts. Additionally, the (substantial) computational resources required for LLM operation raise questions about the environmental sustainability of their widespread adoption in social media platforms. Despite these challenges, the potential benefits of LLMs in enhancing accessibility, improving content quality, and facilitating more meaningful interactions on social media platforms are significant (Raiaan et al., 2023; Hadi, 2023). Given the potential of LLMs in the social media domain as well as the incipient stage of building LLMs-based solutions for understanding conflict narratives in various social media platforms, this research aims to explore their potential application in the context of the ongoing Israel-Palestinian war by focusing on understanding existing war narratives in YouTube. In this sense, this research contributes to a more comprehensive and objective understanding of the conflict, ultimately informing more effective conflict resolution strategies and supporting humanitarian efforts in the region. This is done by within the Design Science Research methodology and considering a Data Science research approach (Peffers et al., 2007) to build a LLMs-based solution. Accordingly, clusters of narratives are captured and knowledge graphs are built in order to reflect on the evolving discourses surrounding the Isrrael-Hamas conflict in an objective and transparent manner (Maathuis & Chockalingam, 2023).

The outline of this article is structured as follows. Section 2 discusses relevant studies characterizing the research perspectives adopted in this research. Section 3 presents the methodological approach considered together with implementation details. Section 4 reflects on the results obtained. And Section 5 provides concluding remarks and future research perspectives.

2. Related Research

Chen et al., (2024) propose the IsmasRed dataset containing public discourses on Reddit regarding the Israel-Hamas conflict from August to November 2023. Comprising nearly 400,000 conversations and over 8 million comments, the dataset demonstrates highly emotionally charged discussions, reflecting the complex and sensitive nature of the conflict. The research points out the intricate interplay between ideology, sentiment, and community engagement in digital spaces, examining trends in topics, controversy, and emotional and moral language over time. Moreover, the study aims to enrich understanding of online discussions about the conflict, providing valuable insights into how people perceive and engage with this contentious issue on social media platforms. Anisa, Harjoko & Khadifa (2023) analyse the role of the #Free Palestine hashtag in the current Israel-Palestinian conflict based on data collected from 06 to 25 December 2023 from X (former Twitter), Facebook, and Instagram. The study highlights the power of cultural communication through social media platforms in transforming public perceptions, influencing policy decisions, and fostering activist movements on a global scale. This is done through a movement of online activism that was further translated into real-world actions that impacted international narratives surrounding various geopolitical and social dimensions of this war. At the same time, the authors analyse the underlying sentiments of the narrative captured in these social media platforms and find that while X served as a platform for polarized political dialogue, with hashtags like #IsraelVsPalestine fostering intense debates Facebook, in contrast, has become a hub for Palestinian solidarity movements, with #FreePalestine garnering significant support and interactions. Instagram's impact is primarily visual, utilizing emotionally charged images from conflict zones to influence public opinion effectively. Focusing on conventional media, Irfan, Aqeel & Hussain (2024) stress out that biased narratives can significantly influence public perception, shape policy decisions, and impact transnational relations, potentially exacerbating tensions and obstructing peace efforts. The findings emphasize the critical role of the media in shaping global perceptions and policy formulation, highlighting the need for media literacy among viewers and the encouragement of journalistic practices that stimulate well-informed and least-biased discussions on complex geopolitical conflicts.

Knowledge graphs are complex data structures that represent information in a graphical format, consisting of interconnected entities and relationships. At their core, knowledge graphs are defined by a set of entities (E), relations (R), factual triples (T), and a function (F) that represents background knowledge. This structure allows for the organization of information in an intuitive and easily navigable manner, facilitating complex queries and data analysis. The graphical representation of knowledge enables the capture of contextual information, which is often challenging to represent in traditional data management systems, making knowledge graphs particularly valuable for handling real-world knowledge that is situational, layered, and dynamic. The primary components of a knowledge graph include entities, relations, factual triples, and background knowledge. Entities represent the main concepts or objects within the graph, while relations define the connections

between these entities. Factual triples form the basic units of information, consisting of two entities linked by a relation. Background knowledge serves as the foundation for constructing the knowledge graph, providing the rules or schema that guide the formation of factual triples and their relationships. The construction process of a knowledge graph involves three key stages: knowledge acquisition, where data is collected and initial structures are formed; knowledge refinement, which improves accuracy and completeness; and knowledge evolution, ensuring the graph remains up-to-date and relevant over time (Zhong et al., 2023). Moreover, knowledge graphs have diverse applications across multiple domains, significantly enhancing the capabilities of various systems and platforms. In search engines, they improve query understanding and result relevance, while in recommendation systems, they enable personalized suggestions by leveraging entity relationships. Question answering systems benefit from the structured information storage and retrieval capabilities of knowledge graphs, allowing for more precise responses to complex queries. In enterprise settings, knowledge graphs facilitate efficient organization and access to internal data, boosting productivity and decision-making processes. Chen et al., (2023)

Moreover, domain-specific applications in fields such as healthcare, finance, and education utilize knowledge graphs to organize specialized knowledge, enabling professionals to access relevant information swiftly and effectively. These wide-ranging applications demonstrate the versatility and power of knowledge graphs in enhancing information retrieval, analysis, and decision-making across various sectors (Zou, 2020; Abu Salih, 2021; Zhong et al., 2023). In the context of using knowledge graphs for summarization and analysis, the focus can then be done on both the textual content and the structural relationships represented in the knowledge graph, resulting in the selection of sentences that most accurately capture the main ideas of the text. The integration of knowledge graphs in this process significantly enhances the quality of the generated summaries, producing more concise and representative extracts that effectively capture the essence of the source document (Gao, 2023). This could be done in a mechanism such as the one integrating the conversational capabilities of AI models with the dynamic information retrieval strengths of QASs (Question Answering Systems). This integration presents a promising direction for developing more powerful and effective knowledge graphs chatbots, capable of combining natural language interaction with accurate, up-to-date information retrieval (Omar et al., 2023).

LLMs are seen as powerful tools in the detection of disinformation and fake news as they contain extensive world knowledge and strong reasoning abilities to detect and counter false information effectively. Nevertheless, LLMs have a dual nature as they can be both powerful tools against social media manipulation mechanisms such as disinformation and misinformation, but they can they can equally be exploited to generate deceptive content at scale, posing a critical challenge. The historical review of misinformation detection methods and current efforts to utilize LLMs for this purpose provide a comprehensive backdrop, emphasizing the need for robust intervention and attribution strategies. A major challenge identified is the increased difficulty in detecting LLM-generated misinformation, which tends to be more deceptive and harmful than human-written content (Chen & Shu, 2023). On the dual nature of LLMs and their potential fairness risks reflect Castillo-Eslava et al., (2023). Accordingly, the authors conduct a series of case studies comparing ChatGPT's responses with Wikipedia information and United Nations resolutions on controversial territories such as Crimea, the West Bank, and Transnistria. Then the paper illustrates AI's potential to consolidate power and shape public understanding of reality. The findings reveal the necessity for ongoing monitoring and analysis of AI's role in constructing legitimacy and recognizing territorial sovereignty, advocating for a critical approach to emerging technologies.

While LLMs like GPT 3.5 demonstrated the ability to identify fake news and provide multi-perspective rationales, in a study conducted by Hu et al., (2024) they unexpectedly underperform compared to simpler, fine-tuned small language models (SLMs) such as BERT. This discrepancy is attributed to LLMs' challenges in effectively selecting and integrating rationales to form conclusive judgments. However, the authors point out that a promising role for LLMs as advisors to SLMs, leveraging their capacity to generate multi-perspective instructive rationales. This suggests a potential synergy between LLMs and SLMs in enhancing fake news detection, where LLMs contribute depth and breadth of analysis, while SLMs excel in focused, task-specific decision-making. Hence, these results reflect the importance of a balanced approach in leveraging different types of language models for optimal performance in combating misinformation. In contrast, in a different context, Sprenkamp, Jones & Zavolokina (2023) analyse the capabilities of GPT-3 and GPT-4 models to identify propaganda within news articles. Tested against a complex dataset known for subtle propaganda techniques and contextual dependencies, GPT-4 demonstrated performance comparable to state-of-the-art models like RoBERTa. The research employed five variations of GPT-3 and GPT-4, applying different prompt engineering

and fine-tuning strategies, to comprehensively evaluate their propaganda detection abilities. Performance was assessed using metrics such as F1 score, Precision, and Recall, providing a robust evaluation of the models' accuracy and reliability. While the study highlights the potential of LLMs in tackling complex tasks like propaganda detection, it also underscores the challenges that remain, emphasizing the need for further research to enhance their effectiveness in real-world applications. These findings contribute valuable insights to the ongoing development of AI-driven tools for combating misinformation and propaganda in digital media. Particularly in the context of the ongoing war in Ukraine, Pavlyshenko (2023) built a fine-tunned LLaMA 2 model using efficient techniques like PEFT/LoRA to analyze a wide range of text data, including news articles, social media posts, and propaganda narratives. The fine-tuning process enables LLMs to perform complex tasks such as fact-checking, identifying manipulation tactics, and detecting fake news by examining semantic structures, frequency of specific item sets, and associative rules within texts. Additionally, LLMs analysed user behaviour on social media platforms to identify patterns indicative of disinformation campaigns Maathuis & Kerkhof (2023b), such as artificial amplification of messages.

While various studies focus on either various approaches and development of LLMs-based solutions for various analytical purposes and knowledge graphs are increasingly considered for implementation in various scientific and practical applications, based on the comprehensive analysis above conducted on the use of LLMs and knowledge graphs in conflict contexts, and especially focusing on the ongoing conflict in Gaza, it can be seen that a limited number of studies are dedicated to this topic while the complexity of the events and their corresponding narratives and underlying sentiments and emotions continue to evolve. It is then the objective of this research to tackle this knowledge gap.

3. Research Approach and Implementation

This research aims to develop and evaluate an innovative LLMs-based solution for detecting and analysing narratives related to the Israel-Hamas conflict within YouTube videos. The proposed system automates the process of video search and narrative detection, initiating from a single search term and iteratively expanding through grouping similar narratives and generating new search terms. This iterative approach facilitates the compilation of a comprehensive dataset of narratives, enabling a multifaceted analysis of the conflict. The solution extends beyond text-based narrative collection by incorporating a knowledge graph visualization, which elucidates the interconnections between identified narratives. This integrated approach combines advanced NLP (Natural Language Processing) techniques with data visualization to provide a robust tool for understanding the complex narrative landscape surrounding the Israel-Hamas conflict, offering potential applications in conflict analysis, media studies, and political science research. To this end, a Design Science Research methodological approach is considered taking a Data Science stance seeing previous successful applications in war context (Peffers et al., 2007; Maathuis, C., & Kerkhof, 2023) and is depicted in Figure 1. Accordingly, the following research question is formulated: How to detect narratives in YouTube video transcripts on the Israel-Hamas conflict using LLMs?

The data collection starts by establishing the search term which is “Israel Hamas”. The term was used to retrieve YouTube videos using the DuckDuckGo Search platform. For each video, transcripts are obtained from YouTube and the videos without a transcript are discarded. Furthermore, the narratives are extracted using GPT-4 and LangChain to facilitate interaction with GPT-4. The prompt objective is defined as follows:

> *Your objective is to extract and summarize the narratives in this text. Extract at most 10 narratives. If there are more, extract the ones most likely to contain disinformation or hate speech.*

As multiple narratives can be extracted from a single video, and the relation between them is stored, so that the reasoning of the LLM can be verified. Then the narratives are clustered with CPT-4, and for each cluster, a unified narrative is created. The prompt objective used in this step is:

> *Cluster these narratives in similar narratives and create a new narrative description for each cluster in max. 100 words each..*

Through a series of experiments and a variation of the number of clusters, the narratives were grouped effectively and the narratives were merged to avoid duplicates in three iterations. Further, triples are extracted from each narrative to create the knowledge graph. The corresponding prompt objective used is depicted below:

Extract triples for this text that can be used in a knowledge graph. For verbs use present tense. Try to prevent complex triple parts consisting of multiple words.'

Standardize these triples by replacing synonyms and alternative phrasings in these triples with a single, consistent text throughout all triples, so that I can create a single consistent knowledge graph with them.

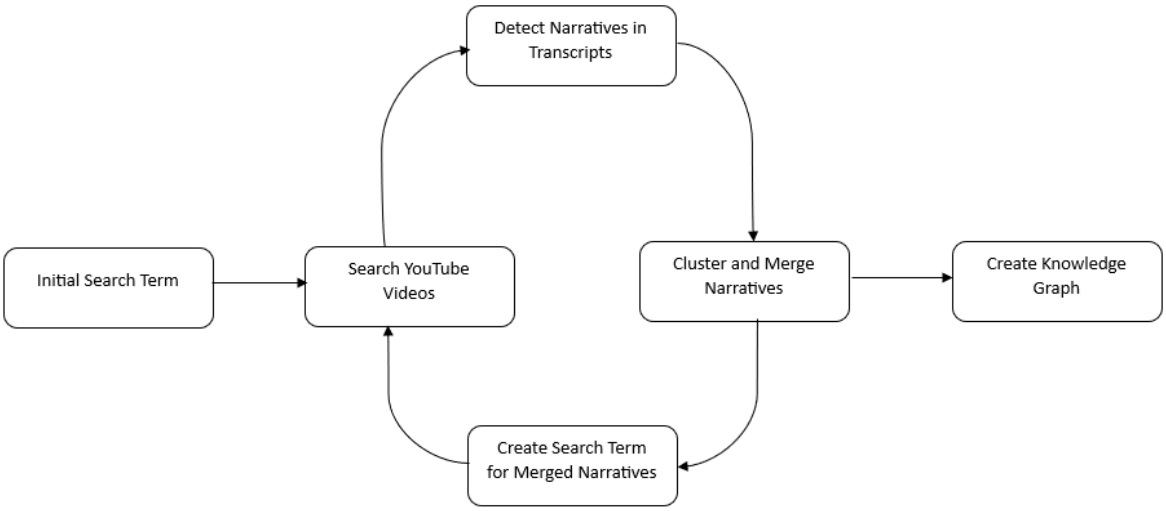


Figure 1: Research approach and implementation

4. Results

A total of 230 videos were found and 61 clusters were initially identified and then merged into 23 narratives. This is captured in Figure 2 and below a selection of three narratives are provided for exemplification and reflection.

Iteration	Narratives	Clusters
1	246	12
2	1077	29
3	882	20
Total	**2205**	**61**

Figure 2: Number of narratives and clusters per iteration

Narrative 1: Hamas and Israel pursue ceasefire negotiations, assess potential prisoner swaps, and discuss phased conflict de-escalation. Proposals include hostage releases and call for the destruction of Hamas to end hostilities, with involvement from international and regional negotiators like the US, Qatar, and Egypt. Reports show varied progress and stances on Israeli withdrawal and a halt in operations.

Narrative 6: Gaza suffers a humanitarian crisis with high casualties, infrastructure damage, and shortages of essentials, eliciting varying international responses and increased international aid efforts.

Narrative 21: Globally, security forces stay alert for conflict linked threats, and educational institutions grapple with internal divisions on the Israel-Hamas issue.

From the final set of narratives, a knowledge graph was constructed by extracting and standardizing the triplets. In this way, 164 triplets were extracted. Accordingly, a selection of triplets is further captured and a part of the knowledge graphs is illustrated in Figure 3 and 4:

(Israel, risks, war crimes prosecution),

(Israel, seeks to, dismantle Hamas),

(Israel, takes, security actions),

(Journalistic disputes, question, authenticity),

(Likelihood, is of, regional war expansion).

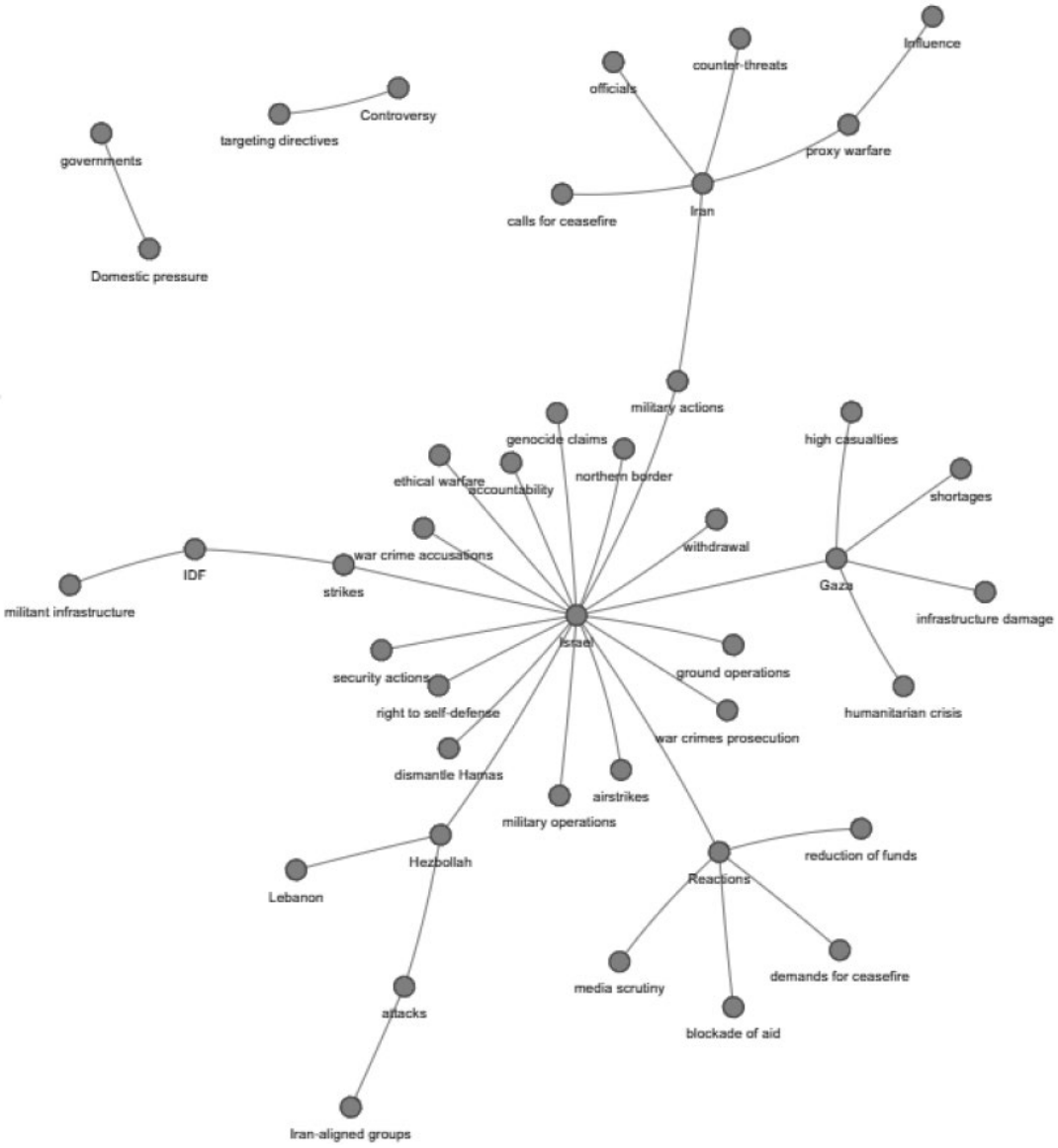


Figure 3: Part of the Knowledge Grap centred around Israel



Figure 4: Part of the Knowledge Grap centred around Hamas

As Figure 3 and Figure 4 show, the ongoing narratives in the Israel-Hamas war, as captured through keywords such as the ones depicted, revolve around several critical themes. These include military actions, humanitarian crises, and high casualties, reflecting the intense and often devastating nature of the conflict. Discussions also highlight security actions, strikes, and proxy warfare, indicating the broader geopolitical implications. There are calls for ceasefires and negotiations on captives, military withdrawal, and prisoner exchanges, underscoring the complex diplomatic efforts involved. The narratives are further complicated by social media manipulation mechanisms like disinformation and misinformation since they are shaping public opinion and influencing international perspectives on the conflict.

The results obtained are analysed in a qualitative way given the approach and context considered in this research. They show that the application of LLMs for narrative analysis has potential to capture important underlying dimensions of the narratives and allows their positioning in relation to ongoing media and research trends and discourses for the ongoing Israel-Hamas war. As lessons learned, LLMs offer enhanced explainability, allowing for the tracing of clustered narratives back to their foundational elements exceling in abstractive clustering, grouping texts based on topical relationships rather than mere word similarity, surpassing traditional machine learning techniques. LLMs also demonstrate improved triple extraction and coreference resolution, and generate coherent narratives instead of hard-to-interpret bags of words common in topic modelling. Nevertheless, the subjectivity in assessing the quality and effectiveness of narrative division poses challenges in quantitative evaluation. Inconsistency is evident in the fluctuating number of clusters and narratives identified across multiple executions, highlighting the process's unpredictability. At the same time, while infrequent, formatting issues, could disrupt the parsing process, necessitating repeated attempts. Perhaps most significantly, the use of advanced LLMs like GPT-4 incurs substantial costs and slower processing speeds, limiting their feasibility for large-scale projects without significant investment.

5. Conclusions

This research proposes an innovative modelling approach to analyse the Israel-Hamas war narratives using LLMs on YouTube video transcripts. To this end, an intelligent solution is developed to iteratively expand from a single search term to detect, group, and visualize narratives through knowledge graphs. In this approach, data visualization methods are used in order to provide a method for visualizing and understanding the complex narratives in the landscape of this conflict. To this end, a Data Science approach is taken in a Design Science Research methodological approach and valuable insights are obtained for further conflict analysis, media studies, and political science research. Future research perspectives may include integrating real-time data streams for dynamic narrative tracking, incorporating multi-modal analysis by combining video and audio content with transcript data, and developing cross-platform comparisons to understand narrative propagation across different social media ecosystems. Additionally, investigating the application of this methodology to other complex geopolitical conflicts could yield valuable comparative insights. Conclusively, this research not only offers a novel approach to understanding the narrative landscape of the Israel-Hamas conflict but also paves the way for building advanced AI-based solutions in conflict analysis and media studies.

References

Abu-Salih, B. (2021). Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications*, *185*, 103076.

Abudayya, A., Bruaset, G. T. F., Nyhus, H. B., Aburukba, R., & Tofthagen, R. (2023). Consequences of war-related traumatic stress among Palestinian young people in the Gaza Strip: A scoping review. *Mental Health & Prevention*, 200305.

Akgül-Açıkmeşe, S., & Özel, S. (2024). EU Policy towards the Israel-Palestine Conflict: The Limitations of Mitigation Strategies. *The International Spectator*, *59*(1), 59-78.

Anisa, N., Harjoko, I., & Khadifa, R. (2023). A Case Study of the Hashtag# Freepalestine: Cultural Communication and Public Opinion Transformation. *Ilomata International Journal of Social Science*, *4*(4), 787-803.

Buheji, M., & Hasan, A. (2024). Beyond Famine and Chaos–Case of Gaza. *International Journal of Management (IJM)*, *15*(2), 2024.

Buheji, M., Mushimiyimana, E., & Ahmed, D. (2024). Empathic engagement with Gaza: dynamics, impact, and prospects. *International Journal of Management (IJM)*, *15*(1), 132-156.

Burghardt, K., Zhang, J., & Lerman, K. (2024). IsamasRed: A Public Dataset Tracking Reddit Discussions on Israel-Hamas Conflict. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 18, pp. 1900-1912).

Castillo-Eslava, F., Mougan, C., Romero-Reche, A., & Staab, S. (2023). The role of large language models in the recognition of territorial sovereignty: An analysis of the construction of legitimacy. *arXiv preprint arXiv:2304.06030*.

Chen, C., & Shu, K. (2023). Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Chen, J., Geng, Y., Chen, Z., Pan, J. Z., He, Y., Zhang, W., ... & Chen, H. (2023). Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. *Proceedings of the IEEE*, *111*(6), 653-685.Chen, K., He, Z.,

Dostri, O. (2024). Israel's Struggle with the Information Dimension and Influence Operations during the Gaza War. *MILITARY REVIEW*, *1*.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Gao, Y. (2023). *A Deep Learning Approach to Extractive Text Summarization Using Knowledge Graph and Language Model* (Master's thesis, Illinois State University).Zhong, L., Wu, J., Li, Q., Peng, H., & Wu, X. A. (2023) comprehensive survey on automatic knowledge graph construction. arXiv 2023. *arXiv preprint arXiv:2302.05019*.

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., & Qi, P. (2024) Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. arXiv 2024. *arXiv preprint arXiv:2309.12247*.

ICC (2024). PANEL OF EXPERTS IN INTERNATIONAL LAW Convened by the Prosecutor of the International Criminal Court, 20 May 2024.

Irfan, I., Aqeel, M., & Hussain, M. S. (2024). Critical Discourse Analysis of Islamophobia Reflection in Gaza-Israel Conflict: A Case Study of BBC Reporting. *Journal of Arts and Linguistics Studies*, *2*(1), 289-301.

Maathuis, C., & Kerkhof, I. (2023). The first two months in the war in Ukraine through topic modeling and sentiment analysis. *Regional Science Policy & Practice*, *15*(1), 56-75.

Maathuis, C., & Kerkhof, I. (2023b). Social media manipulation awareness through deep learning based disinformation generation. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 227-236).

Maathuis, C., & Chockalingam, S. (2023). Modelling Responsible Digital Security Behaviour for Countering Social Media Manipulation. *Indian Institute for Energy Technology*.

Maathuis, C., & Chockalingam, S. (2024). Risk Assessment of Large Language Models Beyond Apocalyptic Visions. In *European Conference on Cyber Warfare and Security* (Vol. 23, No. 1, pp. 279-286).

Myers, D., Mohawesh, R., Chellaboina, V. I., Sathvik, A. L., Venkatesh, P., Ho, Y. H., ... & Jararweh, Y. (2024). Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. *Cluster Computing*, *27*(1), 1-26.

Omar, R., Mangukiya, O., Kalnis, P., & Mansour, E. ChatGPT versus Traditional Question Answering for Knowledge Graphs: Current Status and Future Directions Towards Knowledge Graph Chatbots.(2023) doi: 10.48550. *arXiv preprint ARXIV.2302.06466*.

Pavlyshenko, B. M. (2023). Analysis of disinformation and fake news detection using fine-tuned large language model. *arXiv preprint arXiv:2309.04704*.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.

Pycińska, M. (2023). Israeli and Palestinian settler colonialism in new media: the case of roots. *Humanities*, *12*(5), 124.

Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., ... & Azam, S. (2024). A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Sprenkamp, K., Jones, D. G., & Zavolokina, L. (2023). Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.

Ye, W., Ou, M., Li, T., Ma, X., Yanggong, Y., Wu, S., ... & Zhao, J. (2023). Assessing hidden risks of LLMs: an empirical study on robustness, consistency, and credibility. *arXiv preprint arXiv:2305.10235*.

Zou, X. (2020, March). A survey on application of knowledge graph. In *Journal of Physics: Conference Series* (Vol. 1487, No. 1, p. 012016). IOP Publishing.

Unveiling the Emotional Landscape of Ukrainian War Narratives Using Large Language Models

Clara Maathuis[1] and Iddo Kerkhof[2]
[1]Open University of the Netherlands, Heerlen, The Netherlands
[2]Independent Researcher, The Netherlands

clara.maathuis@ou.nl
email@iddo.eu

Abstract: Social media platforms serve as a dynamic environment where narratives, sentiments, and ideologies intermingle, catalysing the amplification and dissemination of ideas and experiences. As the ongoing endurance war in Ukraine represents a significant inflection point in the contemporary geopolitical landscape, revealing the urgency for comprehensive understanding and analysis of the implications and consequences of the activities therein. While traditional media outlets as we as research and practitioner efforts tackle various social, economic, and political dimensions of this conflict, efforts directed to capturing and reflecting on the nuanced thoughts and emotions of Ukrainian people are still incipient in relation to unconventional social media platforms like TikTok and Telegram. To address this gap, this research develops a novel LLMs (Large Language Models)-based modelling solution that captures, structures, and analyses the discourses characterizing the sentiments and emotions of Ukrainian Telegram users. This is done pursuing a Data Science methodological approach focusing on the insights collected in the first year of war, thus between February 24, 2022, up to February 23, 2023. By harnessing the analytical capabilities of LLMs, this research aims to bridge the gap between conventional understanding and the nuances of human sentiments, thereby advancing comprehension of the multifaceted dimensions of the ongoing war in Ukraine.

Keywords: Russia-Ukraine war, Generative AI, Large Language Models, LLMs, Sentiment analysis.

1. Introduction

"Theory becomes infinitely more difficult as soon as it touches the realm of moral values."

(Carl von Clausewitz)

The ongoing war in Ukraine has emerged as a defining geopolitical crisis of the 21st century, catalysing a shift in global dynamics and international relations. Being carried out beyond traditional battlefield boundaries is having far-reaching political, ideological, social, and economic impacts on a global scale. This implied the re-evaluation of a series of strategic alliances, foreign policies, and domestic programs in response to this evolving crisis. In this sense, the attention of both conventional as well as unconventional media channels (e.g., from social media platforms) was captured, fact that created a multifaceted information landscape and a multi-angled battlefield. In this way, social media platforms serve as both an environment for information dissemination and a chronicler of personal experiences and sentiments. This dual role of social media platforms has enabled a broader and more immediate engagement with the war, influencing public perception and policy responses globally. Moreover, the integration of advanced technologies based on AI, in analysing social media data has further enriched both strategic and public understanding understanding of the conflict's psychological, social, legal, and ethical dimensions, highlighting the complex interplay between modern warfare and digital communication.

At the same time, social media manipulation mechanisms such as disinformation and misinformation, have become intricately entwined with authentic war insights, creating a challenging information landscape (Käihkö , 2020; Fard & Maathuis, 2021; Sufi, 2023). This represents a critical aspect for both practitioner and academic efforts that aim to understand the narratives and sentiments of those directly affected by the conflict, enabling a clearer distinction between truth and falsehood while reflecting the war's broader effects (Alibudbud, 2022; Scharbert et al., 2024). In particular, Telegram emerged as a prominent unconventional social media platform during this conflict, gaining popularity due to its diverse channels and capacity to deliver up-to-date war-related information (DFRLab, 2024; Ptaszek, Yuskiv & Khomych, 2023). Nevertheless, based on the extensive literature review conducted in this research, it can be seen that both research and practitioner efforts dedicated to understanding the dynamics or the narratives and feelings of Telegram users in relation to this war, are still in an incipient phase.

To this end, this research aims to contribute to the ongoing efforts of building intelligent solutions that support and enhance the understanding of the sentiments and emotions experienced in the ongoing war in Ukraine particularly focusing on Ukrainian content. It does that by employing a Data Science methodological approach

(Shearer, 2000; Peffers et al., 2007) and building an innovative LLMs-based modelling solution to process and analyse Telegram data collected from February 24, 2022, to February 23, 2023. This platform was selected based on two arguments: first, its raise in popularity, significance, and use in the last years, and especially during this war; and second, the fact that there is a need of objective analysis of existing narratives discussed in such platforms. By combining LLM-derived insights with traditional analytical techniques, the study offers a comprehensive and multi-faceted perspective on the conflict, encompassing both factual events and the underlying psychological and social dynamics that influence public perception and policy decisions. The goal is to enhance public understanding of the Ukraine war, potentially informing more effective strategies for conflict resolution and humanitarian intervention, and further contribute to ongoing research and practitioner efforts designed to counter various disinformation and misinformation campaigns (Maathuis & Godschalk, 2023) meant to alter or deceive the public opinion in relation to the events, activities, experiences, thoughts, and sentiments experienced in this war.

The outline of this article is structured as follows. Section 2 discusses the background and relates studies to this research. Section 3 presents the methodological and modelling approach considered in order to achieve the objective of this research. Section 4 shows and interprets the results obtained. A the end, in Section 5 concluding remarks and future research perspectives considered are discussed.

2. Related Research

Rodríguez-Ibánez et al., (2023) conduct a comprehensive examination on sentiment analysis in various social media platforms highlighting its significance in diverse domains like finance, health, marketing, politics, and emergency management. The authors stress the importance of joining academic with practitioner efforts when using such methods and point the importance of considering in this analysis or when interpretation its findings, the temporal dynamics and causal relationships between concepts. Al Maruf et al., (2022) build an emotion detection model and examine the meaning and implications of racism in the context of the first months in the ongoing war in Ukraine. For this, the authors use a combination of machine learning, ensemble, and NLP methods applied to Kaggle and Twitter data. The findings show the effectiveness of the boosting approach in categorizing tweets as racist or non-racist, showcasing the model's ability to analyse and identify instances of racism in textual data. Jaradat et al., (2023) presents a hybrid sentiment-discourse analysis model for analysing Facebook posts related to the war in Ukraine in Jordanian dialect. The model combines lexical-based and machine-learning techniques to improve sentiment analysis by incorporating discourse structure and relations. In this way, the authors aim to extract and classify users' opinions and sentiments, particularly focusing on the polarity of the posts. The study collected data from Facebook posts using specific hashtags: a total of 7198 posts and comments in Arabic. The model achieved a high accuracy of 94.68% in detecting sentiment polarity, showcasing the effectiveness of integrating discourse analysis with sentiment analysis in the Jordanian-Arabic context.

Aslan (2023) conducts a Twitter-based topic modelling and sentiment analysis on the war in Ukraine applying various deep learning techniques like CNN (Convolutional Neural Networks) and BILSTM (Bidirectional Long Short-Term Memory). The study aims to provide a comprehensive understanding of people's emotional tendencies and situational awareness during crisis periods, offering a unique data source for unbiased press reports and articles. Along these lines, Maathuis & Kerkhof (2023) analyse the discourses together with their corresponding sentiments in the first months of war in Ukraine focusing on Telegram. As the complexity of this conflict is high implying the use of both conventional and unconventional means and methods, the authors emphasize that the effects of these events impact various aspects of human life and cognition. Focusing on the Reddit platform, Guerra & Karakus (2023) analyse the sentiments of users expressed in relation to the ongoing armed conflict in Ukraine. The study aims to understand the emotional responses in relation to this war, particularly focusing on feelings of hope and fear. Focusing in particular on the Western side, various emotions are captured in respect to aspects like influence diplomatic decisions, aid provisions, and military support to Ukraine, facts that are seen as potentially shaping the course and outcomes of the war. In their Twitter-based sentiment analysis, Garcia & Yabut (2022) find a general negative trend in the sentiments, with sadness being the most prominent emotion. Next to these, the authors consider emotions such as anger, joy, fear, disgust, and surprise. The prevalence of sadness in the tweets indicates a strong emotional response to the conflict, reflecting the sombre and distressing nature of the situation. In addition, sadness as the dominant emotion suggests a collective sense of empathy, sorrow, and concern among individuals engaging with the topic on social media platforms like Twitter.

Taking into account recent AI developments in the area of LLMs, Zhang et al., (2023) explore the application of LLMs like ChatGPT in sentiment analysis tasks, ranging from basic sentiment classification to more complex analyses like aspect-based sentiment analysis and multifaceted subjective text analysis. On this behalf, the study stresses the importance of structured data for better summarization and detailed sentiment analysis. Nevertheless, the article focuses on the importance that evaluation of LLMs plays in respect to specific tasks, but is not considering their application in the context of the Ukrainian war or any other specific conflict. Accordingly, Sirisha et al., (2022) discuss different machine learning and deep learning techniques used in order to analyse the emotions and opinions of people regarding the Russian-Ukrainian war through Twitter data. To this end the authors introduce a novel approach that combines RoBERTa with Aspect Based Sentiment Analysis (ABSA) and LSTM (Long Short-Term Memory) for sentiment analysis, achieving a high accuracy of 94.7%. Along these lines, Ramos & Chang (2023) analyse the sentiments in tweets related to the ongoing war in Ukraine focusing on the first part of the current conflict. The study uses event-related hashtags and DistilRoBERTa variant for the English tweets (603.552) and XLM-RoBERTa-Base model for Russian tweets. The findings obtained reveal that fear and anger were seen as predominant emotions in the English tweets, while the majority of the Russian tweets exhibited negative polarity. Moreover, the authors stress the widespread concern and emotional reactions of individuals towards the ongoing war in Ukraine, pointing out the fact that future research needs to be conducted in order to expand and explore sentiments across different platforms and societal groupings. As topics characterizing the discourses analysed could be mentioned crisis, weapons, and fatalities resulting from the conflict, aspects that indicate the emotional impact and humanitarian aspects of the ongoing crisis. Moreover, Hasan et al., (2023) focuses on sentiment analysis in the Bangla language concerning the Russia-Ukraine war, highlighting the scarcity of high-quality datasets for such tasks in Bangla compared to English. Accordingly, a Bangla annotated dataset was created by collecting and labelling 10.861 Bangla comments from YouTube TV news channels in Bangladesh, classifying them into three sentiment categories: neutral, pro-Ukraine (positive), and pro-Russia (negative). Based on the findings obtained, the minatory class was constituted by the pro-Ukraine class.

As this extensive literature review shows, various ML and DL-based solutions have been proposed with the scope of analysing diverse aspects related to the ongoing war in Ukraine and beyond. Nevertheless, the efforts dedicated to conducted such an in-depth analytical approach using LLMs is incipient, but needed given their advanced natural language understanding and generation capabilities that allow capturing unique insights and contextual aspects. This is the knowledge gap that this research intends to tackle.

3. Research Methodology

To address the knowledge gap identified in this domain, the main objective of this research is to analyse the sentiments and emotions characterising the events from the first year of war in Ukraine on unconventional, yet increasingly used social media platforms like Telegram. In order to do that, a novel LLMs-based modelling solution is developed in a Data Science methodological approach (Shearer, 2000; Peffers et al., 2007). On this behalf, Telegram data is collected between February 24, 2022, up to February 23, 2023. lows. By integrating LLMs derived insights with conventional analytical methods, this study seeks to provide a more holistic and nuanced perspective on the conflict, encompassing not only factual events but also the psychological and social dimensions that shape public perception and policy responses. Ultimately, this research aims to contribute to a deeper, more nuanced understanding of the Ukraine war, potentially informing more effective conflict resolution strategies and humanitarian responses.

With this purpose, a dataset of 47483 messages in Ukrainian language are collected and translated using the Google Translate API (Maathuis & Kerkhof, 2024). The data is further processed and the lemmatization mechanism is applied. Then the sentiment analysis mechanism is implemented using ChatGPT 3.5 Turbo considering the following sentiment classes: devastated (- 1.0), alarmed (- 0.75), fearful (- 0.5), concerned (- 0.25), neutral (0.0), supportive (+ 0.25), hopeful (+ 0.5), optimistic (+ 0.75), and elated (+ 1.0). Furthermore, word clouds are implemented as a graphical representation that depicts the frequency and prominence of the words used by users in order to reflect on the thoughts and emotions they experience in this conflict setting. By aggregating and visualizing the most frequently occurring terms, word clouds enable researchers and analysts to quickly identify dominant themes, emotional states, and key concerns within large volumes of social media data (Giannoulakis & Tsapatsoulis, 2021; Vanstone et al., 2024). The size of each word in the cloud typically corresponds to its frequency or importance, allowing for immediate recognition of prevalent sentiments and topics. This technique is particularly beneficial in war-related contexts as it can reveal shifts in

public opinion, highlight emerging narratives, and capture the emotional zeitgeist of affected populations. In these lines, then the word clouds are analysed and positioned in the ongoing international discourses.

4. Results

In this section the results obtained are presented by depicting the word clouds for each category of the nine categories of sentiment considered, i.e., devastated, alarmed, fearful, concerned, neutral, supportive, hopeful, optimistic, and elated. The word clouds are illustrated in Figure 1 and further discussed and positioned in the ongoing scientific and practitioner discourses.



Fearful



Concerned



Neutral



Supportive



Hopeful



Optimistic

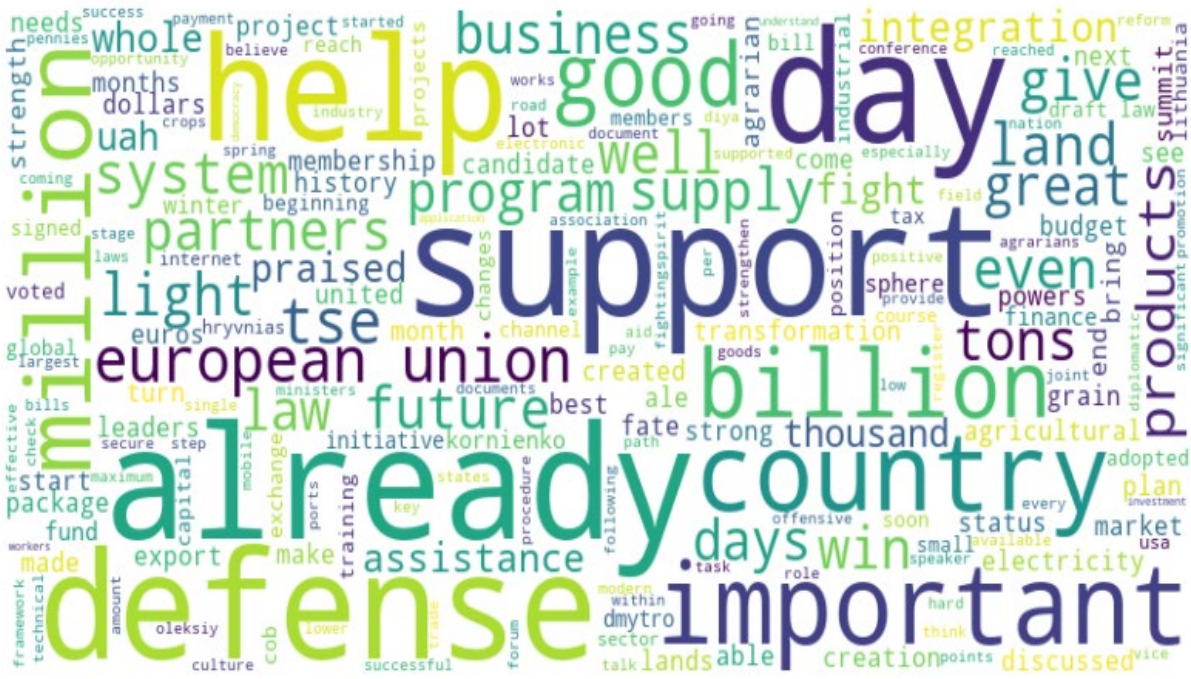




Figure 1: Word clouds of sentiments

In the *devastated* category, words like *aggression, dead, wounded, life, children, territories,* and *result* can be found. These terms point out the severe human and territorial impacts of the conflict, highlighting the widespread perception of violence and loss. "Aggression" and "territories" reflect the geopolitical and military dimensions of the war, while "dead," "wounded," and "life" emphasize the tragic human cost. The inclusion of "children" points to the particular vulnerability and suffering of the youngest victims, and "result" suggests a focus on the outcomes and consequences of the hostilities. This indicates that the public discourse is heavily influenced by the immediate and long-term humanitarian crises resulting from the conflict, painting a grim picture of the war's impact on both individuals and society.

In the *alarmed* category words such as *enemy, alarm, rockets, strikes, artillery, Kiev,* and *night* can be seen. These terms collectively reflect the immediate threats perceived by individuals in relation to the ongoing military actions and their impact on civilian life. "Enemy" and "alarm" reflect the adversarial nature of the conflict and the widespread anxiety it generates. The words "rockets," "strikes," and "artillery" point to specific types of military engagements causing destruction and fear. The inclusion of "Kiev" signifies the focus on the capital city, a central target in the conflict, while "night" suggests the added terror of nocturnal attacks, which exacerbate the sense of vulnerability and helplessness. This shows that the discourse is heavily influenced by the persistent threats posed by the conflict, leading to a pervasive atmosphere of alarm and distress.

In the *fearfu*l category words like *enemy, territory, regions, missiles, troops, settlements,* and *situation* are present. These terms collectively reveal the widespread apprehension regarding the military aspects and territorial implications of the conflict. The presence of "enemy" and "troops" shows the perceived threat from opposing forces, while "territory" and "regions" reflect concerns about geographical control and potential loss of land. "Missiles" emphasizes the fear of long-range attacks and their destructive potential. The inclusion of "settlements" suggests anxiety about the impact on civilian areas, and "situation" indicates a general unease about the evolving state of the conflict. This shows that public discourse is heavily influenced by fear of military escalation, territorial losses, and the overall instability of the war, painting a picture of a population deeply concerned about the immediate and long-term effects of the ongoing hostilities

In the *concerned* category words such *as area, situation, population, infrastructure, defense, Luhansk,* and *Kherson* are seen. These terms collectively show the public's anxiety about the ongoing conflict's impact on specific regions and their inhabitants. "Area" and "situation" reflect general concerns about the geographical and current state of the conflict. The words "population" and "infrastructure" reflect worries about the well-being of civilians and the damage to essential services and facilities. "Defense" indicates apprehension regarding the military capabilities and strategies employed to protect these regions. The inclusion of "Luhansk" and "Kherson" points to specific areas of intense focus and concern, likely due to their strategic importance

and the severity of the conflict there. This demonstrates that public discourse is heavily influenced by concerns over the humanitarian and infrastructural consequences of the war, emphasizing the need for attention to the affected regions and their populations.

In the *neutral* category words like *Kyiv, Lviv, law, rights, budget, system,* and *adopted* can be found. These terms collectively suggest a more detached, analytical perspective on the conflict, emphasizing governance and policy-related issues rather than emotional responses. The inclusion of "Kyiv" and "Lviv" indicates attention to key Ukrainian cities, likely in the context of administrative or strategic discussions. Words like "law," "rights," and "system" point to considerations of legal frameworks and societal structures, while "budget" and "adopted" suggest a focus on financial implications and policy implementations. This shows that a segment of public discourse maintains a relatively objective stance, concentrating on the structural and procedural aspects of the conflict rather than its immediate emotional or humanitarian impacts.

In the *supportive* category words such as *important, help, medical, country, support, assistance,* and *united* can be seen. These terms collectively reveal the public's focus on the necessity and provision of various forms of support to Ukraine. The word "important" highlights the perceived significance of the conflict and the need for intervention. "Help," "support," and "assistance" indicate a widespread call for and recognition of aid efforts, both humanitarian and logistical. The inclusion of "medical" points to specific concerns about healthcare and the provision of medical services to those affected by the war. "Country" and "united" reflect a sense of national and international solidarity, emphasizing the collective effort to support Ukraine. This suggests demonstrates that public discourse is heavily influenced by a sense of duty and cooperation, with a strong focus on providing necessary aid and standing together in support of Ukraine during the conflict.

In the *hopeful* category words like *days, support, good, medical, future, European Union,* and *strength* are present. These terms collectively show a sense of optimism and resilience among the public discourse. The inclusion of "days" suggests a forward-looking perspective, while "support" and "European Union" highlight the importance of international solidarity and assistance. "Good" and "strength" indicate positive sentiments and resilience in the face of adversity. The presence of "medical" points to humanitarian aid efforts, which are often associated with hope for recovery and healing. "Future" emphasizes long-term thinking and aspirations beyond the immediate conflict. This reveals that despite this war, a segment of public discourse maintains a hopeful outlook, focusing on support systems, positive developments, and the potential for a better future, aligning with findings that show varied emotional responses to the conflict across different cultural groups.

In the *optimistic* category words such as *support, billion, important, help, partners, products,* and *summit* are seen. These terms collectively point out to a sense of optimism driven by substantial financial aid and collaborative efforts. The word "support" highlights the ongoing assistance from various entities, while "billion" indicates the significant monetary contributions being pledged or provided. "Important" and "help" reflect the perceived critical nature of these efforts in aiding Ukraine. The inclusion of "partners" and "summit" points to the strategic alliances and high-level meetings that are fostering a united front. "Products" suggests a focus on the economic and material support being mobilized. This shows that public discourse is influenced by a hopeful outlook on the conflict, emphasizing the positive impact of international aid, strategic partnerships, and economic support in bolstering Ukraine's resilience and future prospects.

In the *elated* category words like *medal, youth, sports, athletes, championship, beautiful,* and *love* can be found. These terms collectively reveal a resilient spirit and the continuation of normal life activities, particularly in the realm of sports and youth achievements. The presence of "medal" and "championship" suggests celebrations of athletic accomplishments, while "youth" and "athletes" highlight the role of young people in maintaining a sense of normalcy. "Beautiful" and "love" indicate expressions of positivity and appreciation, potentially as a coping mechanism or a deliberate focus on life's pleasures amidst conflict. This indicates that even in war times, public discourse includes moments of joy and pride, particularly cantered around sports events and youth activities, which may serve as important morale boosters and symbols of national resilience for the Ukrainian people.

This in-depth sentiment analysis reflects a multifaceted emotional landscape, heavily influenced by both immediate and long-term humanitarian crises, significant threats, and fears of military escalation and territorial losses (Kunertova & Masuhr, 2022; Zasiekin et all, 2022; NATO, 2023; Tong, 2024; Kušen & Strembeck, 2024; UN, 2024) . Moreover, an atmosphere of alarm and distress is seen, with significant concern over the humanitarian and infrastructural effects of the war. Despite these challenges, a segment of the discourse maintains a relatively objective stance, focusing on structural and procedural aspects of the conflict. At the same time, there is a strong sense of duty and cooperation, emphasizing the provision of aid and

support for Ukraine (European Commission, 2023; NATO, 2023; Jaradat, 2023; NATO, 2024). Notably, a hopeful outlook persists among some, driven by international aid, strategic partnerships, and economic support, which bolster Ukraine's resilience and future prospects. Additionally, moments of joy and pride, particularly cantered around sports events and youth activities, serve as important morale boosters and symbols of national resilience. This complex emotional response reflects the varied and dynamic nature of the sentiments of Ukrainian people in the face of ongoing hostilities.

5. Conclusions

Social media platforms are dynamic environments where narratives, sentiments, and ideologies intermingle, catalysing the amplification and dissemination of ideas and experiences. The ongoing endurance war in Ukraine represents a significant inflection point in the contemporary geopolitical landscape, underscoring the urgency for a comprehensive understanding and analysis of its implications and consequences. While traditional media outlets and research efforts address various social, economic, and political dimensions of this conflict, efforts to capture and reflect on the nuanced thoughts and emotions of the Ukrainian people remain nascent, particularly concerning unconventional social media platforms like TikTok and Telegram. To address this gap, this research develops a LLMs-based modelling solution to capture and analyse the discourses characterizing the sentiments and emotions of Ukrainian Telegram users. This is achieved through a Data Science methodological approach, focusing on insights collected during the first year of the war, from February 24, 2022, to February 23, 2023. Then, given the clusters of sentiments obtained, this research analyses each of the narratives and positions them in the corresponding war situation in a structured and objective manner. To this end, this research to research and practitioner efforts aimed at using advanced AI technologies and through them, intelligent systems that provide support to the comprehension of the multifaceted dimensions of the ongoing war in Ukraine together with the impact on humans and society. At the same time, the results obtained in this research depict a clear and objective picture of the variety of emotions experienced in this conflict, and have the potential to contribute to ongoing efforts on building conflict resolution strategies and humanitarian responses on this behalf.

While acknowledging the complexity and gravity of this war on the one side, and the significant potential of advanced AI technologies such as the ones based on LLMs, this research continues by designing and building responsible and trustworthy intelligent solutions that analyse this conflict based on multiple sources and various dimensions (Maathuis & Chockalingam, 2023). At the same time, this research presents valuable input that can be considered when building defence modelling and simulation solutions that are meant to understand, analyse, and predict the dimensions, nature, and scale of the impact produced by the strategies taken and events occurring during complex (armed) conflicts.

References

Al Maruf, A., Ziyad, Z. M., Haque, M. M., & Khanam, F. (2022). Emotion detection from text and sentiment analysis of Ukraine Russia war using machine learning technique. *International Journal of Advanced Computer Science and Applications*, *13*(12).

Alibudbud, R. (2022). A web-based analysis of public concern about Post-Traumatic Stress Disorder (PTSD) Ukraine during the 2022 military conflict: An infodemiological study. European Journal of Trauma & Dissociation, vol. 6, no. 3, 100286.

Aslan, S. (2023). A deep learning-based sentiment analysis approach (MF-CNN-BILSTM) and topic modeling of tweets related to the Ukraine–Russia conflict. *Applied Soft Computing*, *143*, 110404.

DRFLab (2024). Another battlefield: Telegram as a digital front in Russia's war against Ukraine.

European Commission (2023). Commission Staff Working Document Ukraine 2023 Report.

Garcia, M. B., & Cunanan-Yabut, A. (2022). Public sentiment and emotion analyses of Twitter data on the 2022 Russian invasion of Ukraine. In *2022 9th international conference on information technology, computer, and electrical engineering (ICITACEE)* (pp. 242-247). IEEE.

Giannoulakis, S., & Tsapatsoulis, N. (2021). Topic Identification via Human Interpretation of Word Clouds: The Case of Instagram Hashtags. In *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 International Conference, AIAI 2021, Hersonissos, Crete, Greece, June 25–27, 2021, Proceedings 17* (pp. 283-294). Springer International Publishing.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Guerra, A., & Karakuş, O. (2023). Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict. *Frontiers in Artificial Intelligence*, *6*, 1163577.

Hasan, M., Islam, L., Jahan, I., Meem, S. M., & Rahman, R. M. (2023). Natural language processing and sentiment analysis on bangla social media comments on Russia–Ukraine war using transformers. *Vietnam Journal of Computer Science*, *10*(03), 329-356.

Jaradat, R., Jaradat, G. M., Alsmadi, M., Alzaqebah, M., Almarashdeh, I., Althunibat, A., & Al-Tarawneh, H. (2023). A Hybrid Sentiment-Discourse Analysis Model for Ukraine Crisis Facebook Posts with a Jordanian Dialect. *International Journal of Advances in Soft Computing & Its Applications*, *15*(2).

Käihkö, I. (2020). Conflict chatnography: Instant messaging apps, social media and conflict ethnography in Ukraine. *Ethnography*, *21*(1), 71-91.

Kunertova, D., & Masuhr, N. (2022). The War against Ukraine Shapes NATO's Future. *CSS Policy Perspectives*, *10*(4).

Kušen, E., & Strembeck, M. (2023). The Effects of Multiple Exposure to Highly Emotional Social Media Content During the Early Stages of the 2022 War in Ukraine. *SN Computer Science*, *4*(5), 663.

Maathuis, C., & Kerkhof, I. (2023). The first two months in the war in Ukraine through topic modeling and sentiment analysis. *Regional Science Policy & Practice*, *15*(1), 56-74.

Maathuis, C., & Godschalk, R. (2023). Social Media Manipulation Deep Learning based Disinformation Detection. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 237-245).

Maathuis, C., & Chockalingam, S. (2023). Modelling Responsible Digital Security Behaviour for Countering Social Media Manipulation. *Indian Institute for Energy Technology*.

Maathuis, C., & Kerkhof, I. (2024). Ukrainian Thoughts and Feelings based on One Year War Content Analysis. In *International Conference on Cyber Warfare and Security* (Vol. 19, No. 1, pp. 137-148).

NATO (2023). Russian War Against Ukraine: Lessons Learned Curriculum Guide.

NATO (2024). NATO Secretary General: "the longer we commit, the sooner Ukraine can have peace", https://www.nato.int/cps/en/natohq/news_226436.htm

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, vol. *24*, no. 3, pp. 45-77. Ptaszek, G., Yuskiv, B., & Khomych, S. (2024). War on frames: Text mining of conflict in Russian and Ukrainian news agency coverage on Telegram during the Russian invasion of Ukraine in 2022. *Media, War & Conflict*, *17*(1), 41-61.

Ramos, L., & Chang, O. (2023). Sentiment analysis of Russia-Ukraine conflict tweets using RoBERTa. *Uniciencia*, *37*(1), 421-431.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, 119862.

Scharbert, J., Humberg, S., Kroencke, L., Reiter, T., Sakel, S., Ter Horst, J., ... & Back, M. D. (2024). Psychological well-being in Europe after the outbreak of war in Ukraine. *Nature Communications*, *15*(1), 1202.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, pp. 13-22.

Sirisha, U., & Bolem, S. C. (2022). Aspect based sentiment & emotion analysis with ROBERTa, LSTM. *International Journal of Advanced Computer Science and Applications*, *13*(11).

Sufi, F. (2023). Social Media Analytics on Russia–Ukraine Cyber War with Natural Language Processing: Perspectives and Challenges. *Information*, *14*(9), 485.

Tong, E. (2024). Repercussions of the Russia–Ukraine war. *International Review of Economics & Finance*, *89*, 366-390.

United Nations (2024). Protection of Civilians in Armed Conflict – March 2024, https://ukraine.un.org/en/265434-protection-civilians-armed-conflict-%E2%80%94-march-2024

Vanstone, M., Toledo, F., Clarke, F., Boyle, A., Giacomini, M., Swinton, M., ... & Cook, D. (2024). Narrative medicine and death in the ICU: word clouds as a visual legacy. *BMJ supportive & palliative care*, *14*(e1), e634-e641.

Zasiekin, S., Kuperman, V., Hlova, I., & Zasiekina, L. (2022). War stories in social media: Personal experience of Russia-Ukraine war. *East European Journal of Psycholinguistics*, *9*(2), 160-170.

Zhang, W., Deng, Y., Liu, B., Pan, S. J., & Bing, L. (2023). Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

AI, Personalized Education, and Challenges

Aziz Mimoudi
Institute of Education Science, Um6P, Rabat, Morocco

Aziz.Mimoudi@Um6p.ma

Abstract: Artificial Intelligence (AI) is gaining traction in education, with potential applications that range from personalized learning to automated administrative tasks. However, the integration of AI into educational systems is not without its challenges. This paper explores both the opportunities and obstacles that AI presents in the field of education, particularly through the use of Intelligent Tutoring Systems (ITS) and Adaptive Learning Management Systems (ALMS). These technologies aim to tailor learning experiences to individual students by analyzing data and adjusting content to suit their needs. While this personalized approach could enhance student engagement and comprehension, it relies on vast amounts of data, raising concerns about privacy and the potential misuse of personal information. Moreover, AI's impact on education extends to supporting educators by providing insights into student performance and automating routine tasks. However, the effectiveness of AI systems in this regard remains questionable, particularly when considering the limitations in current AI models and the challenges of integrating them into existing educational frameworks. The risk of algorithmic bias is also a critical issue, as AI systems can inadvertently reinforce inequalities present in the data they are trained on, leading to unfair or discriminatory outcomes. Additionally, while AI promises to streamline certain aspects of education, there are concerns that over-reliance on technology could depersonalize the learning process. Human educators play a crucial role in fostering not only intellectual growth but also emotional and social development—elements that AI systems are currently unable to replicate. This paper argues that while AI holds significant promise in education, its deployment must be carefully managed, with attention to ethical considerations, equity in access, and the preservation of the human elements essential to effective learning. Addressing these challenges is key to ensuring that AI contributes meaningfully to education rather than exacerbating existing issues.

Keywords: Artificial intelligence in education, Personalized learning, Intelligent tutoring systems, Algorithmic bias, Ethical concerns in AI

1. Introduction

Artificial Intelligence (AI) is increasingly viewed as a disruptive force in education, but its rapid adoption raises as many concerns as it does possibilities. While AI promises to personalize learning and streamline educational processes, the reality of integrating such technology is far more complex. Despite advances in AI, its application in education is still limited by technological, ethical, and logistical challenges.

Personalized learning, often hailed as one of AI's greatest contributions to education, relies heavily on data-driven systems that adapt educational content based on students' interactions. However, this level of individualization comes with a cost. The reliance on vast data collection poses risks to student privacy, while the potential for algorithmic bias threatens to perpetuate existing inequalities in educational outcomes. Intelligent Tutoring Systems (ITS) and Adaptive Learning Management Systems (ALMS) can offer tailored feedback and learning paths, but their effectiveness depends on accurate data and careful design—both of which are often lacking.

Moreover, while AI tools may reduce the administrative burden on educators, there are valid concerns that these technologies could depersonalize the learning experience by reducing the role of human educators. AI systems cannot replicate the nuanced understanding and emotional support that educators provide. This paper aims to provide a more critical analysis of AI's role in education, emphasizing not only its potential benefits but also the significant hurdles that must be addressed to ensure its responsible and equitable use.

2. Personalized Learning with AI

2.1 Intelligent Tutoring Systems and Adaptive Learning Management Systems

AI algorithms have revolutionized the field of education by enabling the creation of tailored educational content through the analysis of vast amounts of data. This process involves utilizing machine learning techniques to identify patterns, trends, and insights within educational datasets, which can then be used to personalize learning experiences for individual students. By leveraging AI algorithms, educators can deliver content that is specifically tailored to meet the needs, and preferences of each learner.

One of the primary applications of AI algorithms in education is in the development of adaptive learning management systems (ALMS) based on Intelligent tutoring systems (ITSs). These systems use AI to analyze data on students' past performance, interactions with learning materials, and other relevant variables to

dynamically adjust the pace, difficulty, and content of lessons (Raj & Renumol 2022). For example, if a student is struggling with a particular concept, the ITS can provide additional support and resources to help them master the material. Conversely, if a student demonstrates mastery of a topic, the system can automatically advance them to more challenging content, ensuring that each learner is appropriately challenged and engaged.

ITS provide personalized tutoring and feedback to students based on their individual strengths, weaknesses, and learning goals. By analyzing data on students' performance in real-time, ITS can offer immediate feedback on multiple-choice questions, quizzes, or interactive activities, helping students gauge their understanding and track their progress in real-time. ITS can also identify areas for improvement, and recommend additional resources or practice exercises. This personalized approach to tutoring fosters a supportive learning environment that empowers students to navigate complex subjects with confidence and competence. ITS could be the largest category of artificial intelligence in education (AIED) applications. Several meta-analytic studies have demonstrated the effectiveness of ITSs to promote learning outcomes (VanLehn 2011).

For example, if a student demonstrates proficiency in a particular topic, the system may advance him to more challenging exercises or higher-level content. Conversely, if a student struggles with a concept, the system can provide additional explanations, examples, or practice opportunities to reinforce understanding. Another example, in language learning applications, AI tutors can analyze students' written responses for grammar errors, vocabulary usage, and coherence, providing corrective feedback and suggestions for improvement. Moreover, ITS can help to scaffold students learning process and promote deeper understanding. By analyzing students' interactions with the system, including their responses to questions, problem-solving strategies, and misconceptions, ITS can identify areas where additional support is needed and provide targeted interventions accordingly (Baker & Inventado 2016).

2.2 Intelligent Tutoring System: A History of Evolution

Intelligent Tutoring Systems (ITSs) have evolved significantly since the seminal work of Benjamin Bloom, who identified the "Two Sigma Problem" highlighting the vast disparity in learning outcomes between one-on-one tutoring and traditional classroom instruction (Bloom 1984). This problem spurred the development of ITSs, aiming to provide personalized and effective learning experiences akin to individual tutoring. This text explores the journey of ITSs from its inception to the integration of cognitive theories, focusing on the contributions of pioneers like John Anderson and the development of the ACT theory (Anderson 1993).

Benjamin Bloom's research in the 1980s revealed that students who received one-on-one tutoring performed two standard deviations better than those in traditional classroom settings. This stark contrast led to the identification of the Two sigma problem, highlighting the potential of personalized instruction in enhancing learning outcomes.

In response to Bloom's findings, researchers began developing ITSs in the 1970s and 1980s. These systems utilized computer technology to provide personalized instruction tailored to individual learner needs. Early examples include the pioneering work of researchers like John Anderson, whose work on the Adaptive Control of Thought (ACT) theory revolutionized the field. Anderson's ACT theory posits that intelligent behavior emerges from the interaction between procedural and declarative knowledge. This theoretical framework provided the foundation for the development of cognitive tutors capable of adapting to individual learner cognitive processes.

This approach has been successfully applied across various domains, including mathematics, science, and language learning (Anderson 1993).

2.3 Architecture of an Intelligent Tutoring System

ITSs encompasses various components and frameworks (VanLehn 2011). aimed at facilitating effective interaction between learners and tutors:

- The learner profile: It captures and represents information about the learner's knowledge, skills, preferences, and learning progress. It enables the system to adapt instructional content and feedback to meet the individual needs of each student. It also encompassing domain-specific knowledge, instructional strategies, and pedagogical content. This component is often represented using knowledge representation languages such as production rules, semantic networks, or ontologies.
- The domain model: Hosts the actual educational content. It notably includes the content management system that stores educational resources such as texts, videos, quizzes, etc. This module

plays a central role in the creation, updating, and distribution of pedagogical content. It functions as a kind of knowledge repository, embodying the specific knowledge of a given domain, such as mathematics, sciences, languages, by containing the necessary rules, concepts, and relationships. Additionally, it contributes to the generation of questions, answers, and the assessment of the learner's understanding.

- The expert model: It enables the ITS to generate and evaluate exercises, assignments, and assessments tailored to the learner's proficiency level and learning objectives. It often involves algorithms for problem generation, solution evaluation, and adaptive task sequencing. This component defines the instructional strategies employed by the ITS to deliver content, provide feedback, and support learning activities. It encompasses techniques such as mastery learning, scaffolding, and problem-solving guidance.
- The communication Interface: It facilitates interaction between the learner and the ITS, providing user-friendly interfaces for accessing instructional materials, receiving feedback, and engaging in learning activities. It may include graphical user interfaces, natural language interfaces, or multimodal interfaces.

However, challenges persist in designing ITS architectures that are flexible, scalable, and capable of supporting diverse learning contexts and domains.

The figure below represents an ITS architecture that we adapted.

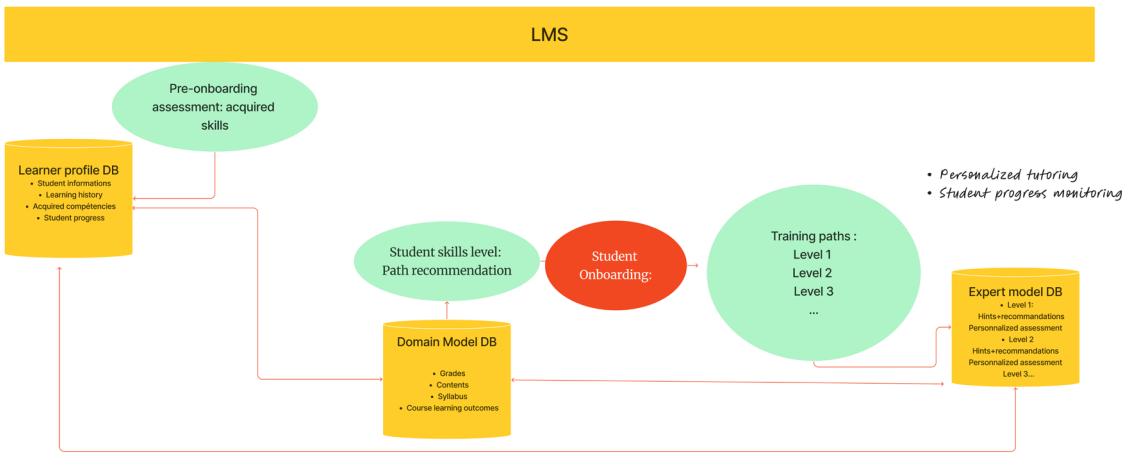


Figure 1: Adapted ITS architecture

Furthermore, AI algorithms can be used to generate educational content that is tailored to meet the specific needs of individual learners. Content AI-generators leverage natural language processing (NLP) and machine learning techniques to create customized learning materials, including textbooks, quizzes, and interactive exercises. These AI-generated resources can adapt to students' preferences, providing a more engaging learning experience.

3. AI Transformers

3.1 Foundation Models, Deep Learning, Machine Learning

Foundation models constitute large-scale AI models that are pre-trained on vast amounts of general data and that can be adapted for downstream applications.

They are at the forefront of generative artificial intelligence, driving advancements in natural language generation, image synthesis, and other creative tasks. These models, such as OpenAI's Generative Pre-Trained Transformer series (GPT) and Google's Bidirectional Encoder Representations from Transformers (BERT), are trained on massive datasets using unsupervised learning techniques to learn rich representations of language and context.

GenAI, a subset of AI and Deep Learning (DL), has surged in recent years, focusing on crafting fresh content like images and text through trained algorithms. In essence, it builds upon AI's groundwork, with Machine Learning facilitating learning from data and DL housing intricate models like neural networks. GenAI, a DL sub-branch,

surpasses mere prediction and classification, employing neural networks to generate entirely novel content across diverse domains (Carlson et al. 2024).

GenAI involves AI systems adept at crafting original content such as text, images, or music, distinct from existing data. Powered by foundation models, it harnesses the core architecture and knowledge necessary for producing coherent and contextually relevant output.

In the dynamic landscape of artificial intelligence (AI), the symbiotic relationship between industry and academia plays a pivotal role in driving innovation and advancement. However, recent trends underscore a notable shift in the balance of power, particularly in the realm of building and disseminating foundational models. While academia has long been regarded as a bastion of knowledge and research in AI, the industry now stands as the dominant force, wielding considerable influence in model development and deployment (Eynon 2024). The latest insights shed light on this evolving narrative, revealing a clear disparity between industry and academia in the realm of AI innovation.

Over the past year, industry titan Google has emerged as a frontrunner, surpassing its competitors in the release of foundational models, including noteworthy additions like Gemini and RT-2. while OpenAI follows closely behind.

In contrast, academia finds itself trailing behind its industry counterparts, grappling with the challenges of keeping pace with the rapid advancements in AI.

Generative Artificial Intelligence (GenAI) can generate new content autonomously, mimicking human-like creativity and problem-solving capabilities. In the context of education, GenAI holds immense potential to revolutionize content creation, offering innovative solutions for personalized and interactive learning experiences. It has the capacity to generate instructional materials, interactive simulations, and virtual environments tailored to individual learner preferences and needs. It can analyze complex educational concepts and generate explanations, examples, and exercises in a manner that is accessible and engaging for learners.

By automating the creation of educational content, GenAI can alleviate the burden on educators and instructional designers, allowing them to focus on pedagogical innovation and the cultivation of critical thinking skills among learners (Eynon 2024).

3.2 Generative AI in Education

Here are some real-world examples of how GenAI is being used for content creation in education:

OpenAI's GPT in Writing Assistance: GPT models are being utilized to assist students and educators in writing tasks. It can generate essays, summaries, and creative pieces based on prompts provided by users (Carlson et al. 2024).

IBM Watson's content creation for training modules: A cognitive computing platform that offers tools for generating interactive training modules and course materials. Educators can use Watson's natural language processing capabilities to analyze course content and automatically generate quizzes, tutorials, and interactive exercises tailored to individual learning objectives and preferences (Chow et al. 2023).

Language learning apps: In his literature review, Law (2024) emphasizes the significance of GenAI tools, such as ChatGPT, in language education, highlighting their potential benefits for both learners and teachers. These technologies are recommended for integration into language lessons to enhance educational outcomes. However, the study identifies significant gaps in current research, particularly the need for more studies beyond the realm of English language teaching. Furthermore, there is a noticeable lack of empirical evidence regarding the effectiveness of GenAI tools in improving educational outcomes across various languages.

Mathematics teaching: GenAI in mathematics teaching, as exemplified by platforms like Wolfram Alpha and Newton, presents both significant opportunities and challenges. Wolfram Alpha, a computational knowledge engine, uses GenAI to create step-by-step solutions for mathematical problems. Students can input equations and receive detailed explanations, aiding in the understanding of complex concepts and improving problem-solving skills. Similarly, Newton employs GenAI to generate personalized learning pathways by analyzing students' learning profiles and performance data, thereby tailoring educational experiences to individual needs.

However, the use of GenAI in math education is not without its criticisms. One major concern is the lack of justification and reasoning provided by AI-generated solutions. In mathematics, explaining why a particular solution is correct is just as important as finding the solution itself. GenAI platforms often fall short in this area, offering answers without sufficient explanation of the underlying principles or the rationale behind the steps taken. This gap is problematic, as research emphasizes the importance of justification in mathematical learning, helping students to develop a deeper understanding and ability to reason through problems.

Furthermore, the AI-generated responses often fail to explicitly identify and correctly use the specialized terminology and concepts of the field. For example, terms essential to the understanding and solving of mathematical problems might be glossed over or used imprecisely, leading to potential confusion. This lack of clarity can hinder students' ability to grasp essential concepts, as they are not provided with the necessary definitions or context that would aid in their comprehension and application.

In addition, while platforms like Wolfram Alpha and Newton aim to enhance learning through personalization, there is concern about the correctness and reliability of AI-generated solutions. Studies have shown that AI-driven chatbots, including those based on large language models (LLMs) like GPT-3.5, GPT-4, and PaLM-2, can sometimes produce inaccurate or misleading solutions to mathematical problems. This issue raises questions about the dependability of these tools in educational settings, particularly when students may rely heavily on them without sufficient critical evaluation ().

Virtual laboratories: GenAI is also used to create interactive learning simulations and virtual laboratories for STEM education. Companies like Labster and PhET Interactive Simulations develop virtual experiments and simulations using GenAI algorithms, allowing students to conduct hands-on experiments in virtual environments and explore scientific concepts in a safe and immersive way.

Interactive storytelling: GenAI algorithms are being used to create interactive storytelling experiences and educational games that engage students in immersive narratives and decision-making scenarios. Platforms like Twine and Ink generate branching storylines and dialogue options based on user input, allowing students to explore different outcomes and perspectives in a dynamic storytelling environment.

Automated essay grading: Platforms like Gradescope and Turnitin utilize GenAI algorithms to automate the grading of essays and written assignments. These systems can analyze the structure, coherence, and language proficiency of student essays, providing instant feedback and assessment to both students and educators (Anon 2024).

These examples illustrate how GenAI is being leveraged to create engaging and personalized learning experiences across various educational domains, from writing assistance and language learning to math problem-solving and interactive simulations.

3.3 GPT's functioning as Intelligent Tutoring System

According to Carmon et al. (2023), GPTs applied to teaching and learning have the potential to function as intelligent tutoring systems. GPT-4 was developed based on machine learning principles (OpenAI et al. 2024), while other well-known intelligent tutoring systems are founded on different theoretical bases. For instance, Cognitive Tutor (Anderson 1993) was developed using the "adaptive control of thought-rational" (ACT-R) theory of cognition. The ACT-R theory outlines a cognitive architecture comprising four modules: perceptual-motor, goal, declarative, and procedural (Anderson 1993). Another theoretical foundation for intelligent tutoring systems is constraint-based modeling, which employs parameters called "constraints" to encode domain knowledge.

Unlike traditional ITSs such as Cognitive Tutor or those based on constraint-based modeling like SQL-Tutor. Anatomy GPT was not designed according to a specific learning theory. However, the introduction of Khanmigo, described by Khan Academy as an "AI-powered teaching assistant" highlights the potential of GPTs to become essential learning tools for modern students at all levels. Combining the machine learning principles underlying GenAI applications with the psychological principles behind traditional intelligent tutoring systems may lead to modern applications that embody the attributes envisioned by the inventors of ITSs.

3.4 Augmenting Educators' Role with AI

AI tools have become invaluable assets in education by providing educators with valuable insights into student progress and learning patterns. These tools utilize sophisticated algorithms to analyze vast amounts of data

generated during the learning process, offering educators actionable information to support personalized instruction and improve learning outcomes.

One of the primary functions of AI tools in education is to track and monitor student progress over time. By collecting and analyzing data from various sources, such as learning management systems, online assessments, and digital learning platforms, AI tools can create comprehensive profiles of each student's academic performance and engagement levels. For example, AI-powered analytics dashboards can visualize students' performance trends, identify areas of strength and weakness, and highlight patterns of behavior that may impact learning outcomes.

Moreover, AI tools can predict students' future performance based on their past behavior and learning history. Machine learning algorithms can analyze historical data to identify patterns and correlations between students' actions, such as study habits, interaction patterns, and assessment scores, and their eventual academic outcomes. By leveraging predictive analytics, educators can anticipate students' needs, identify at-risk learners, and intervene proactively to provide targeted support and interventions (Baker & Inventado 2016).

Furthermore, AI tools enable educators to gain deeper insights into students' learning preferences and cognitive processes. By analyzing students' interactions with digital learning materials, AI algorithms can identify individual learning styles, preferences, and misconceptions, allowing educators to tailor instruction to meet students' unique needs. For instance, adaptive learning platforms can dynamically adjust the pace, difficulty, and content of instruction based on students' responses, ensuring optimal engagement and comprehension (VanLehn 2011).

In addition to providing insights into student progress, AI tools can also support educators in their instructional planning and decision-making processes. Natural language processing algorithms can analyze student-generated content, such as essays, discussions, or projects, to assess comprehension, critical thinking skills, and creativity. By automating routine tasks such as grading and feedback generation, AI tools free up educators' time to focus on higher-order teaching activities, such as designing engaging learning experiences and facilitating class discussions.

Overall, AI tools play a crucial role in enhancing educators' understanding of student progress and learning patterns. By leveraging advanced analytics and machine learning techniques, these tools provide educators with actionable insights to support personalized instruction, identify at-risk learners, and improve learning outcomes in diverse educational settings.

Also, by automating routine administrative tasks, AI enables educators to allocate more time and resources to higher-order teaching activities that require human expertise and creativity. Freed from the burden of manual administrative duties, educators can focus on facilitating critical thinking, problem-solving, and collaboration among students (Graesser et al. 2017).

4. Analysis of Challenges Faced by Educators

Integrating adaptive Learning Management Systems (ALMSs) presents a myriad of challenges for educators, particularly concerning time allocation for application. While the benefits of ALMSs are evident in their ability to personalize learning experiences and improve student outcomes, several obstacles hinder their seamless integration into educational settings.

One of the primary challenges is the steep learning curve associated with adopting new technology. Educators often lack the necessary training and technical expertise to effectively leverage ALMSs in their teaching practices. Without adequate professional development opportunities and support resources, educators may feel overwhelmed by the complexity of the technology and struggle to incorporate it into their instructional strategies.

Furthermore, the time constraints imposed by existing workload demands pose a significant barrier to the implementation of ALMSs. Educators are already stretched thin with their teaching responsibilities, administrative duties, and research commitments. Integrating new technology requires additional time and effort for planning, preparation, and ongoing maintenance, which may exceed educators' available resources.

Moreover, resistance to change among faculty members can impede the adoption of ALMSs. Educators may be hesitant to embrace unfamiliar technology due to fear of failure, concerns about job security, or skepticism

about its efficacy. Overcoming this resistance requires a supportive organizational culture that encourages experimentation, risk-taking, and continuous professional growth.

Addressing these challenges requires a multifaceted approach that combines comprehensive training programs, institutional support structures, and collaborative partnerships between educators, administrators, and technology providers. Providing educators with ongoing professional development opportunities, access to technical support, and incentives for innovation can empower them to overcome obstacles and maximize the potential of ALMSs in enhancing teaching and learning outcomes.

5. Ai and Ethical Considerations

The integration of AI in educational environments has opened new avenues for personalized learning and efficient administrative processes. However, it also raises significant concerns regarding data privacy, algorithmic bias, and the potential depersonalization of learning experiences.

5.1 Data Privacy

One of the primary concerns with the use of AI in education is data privacy. AI systems rely on vast amounts of data to function effectively, including sensitive information about students' academic performance, behavior, and even personal details. This data collection often occurs without explicit consent from students or their guardians, raising ethical questions about transparency and consent.

Ensuring the protection of this data is paramount. The misuse or breach of student data can lead to severe consequences, including identity theft and unauthorized profiling. According to a study by Oneil (2016), educational institutions must adopt robust data governance frameworks to safeguard students' information and maintain trust in AI systems.

5.2 Algorithmic Bias and Inequalities

Algorithmic bias is another critical issue associated with AI in education. AI systems are only as unbiased as the data they are trained on. If the training data contains biases, these biases can be perpetuated and even amplified by the AI system. This is particularly concerning in educational settings, where biased algorithms can affect student outcomes and perpetuate inequalities. Research by Oneil (2016) highlights how biases in search algorithms can reinforce stereotypes and disadvantage marginalized groups.

In the context of AI in education, biased algorithms can manifest in various ways, such as biased grading systems, unequal access to learning resources, and unfair disciplinary measures. For example, an AI-based grading system might inadvertently favor students from certain backgrounds if the training data predominantly reflects the performance patterns of those students. Addressing algorithmic bias requires a multifaceted approach, including diverse data collection, regular auditing of AI systems, and the inclusion of ethicists in the development process to ensure fairness and equity (Eynon 2024).

5.3 Depersonalization of Learning Processes

The depersonalization of learning processes is another significant concern with the adoption of AI in education. While AI can offer personalized learning pathways, there is a risk that it may reduce the human element that is crucial to effective teaching and learning. The interpersonal relationship between educators and students is a fundamental aspect of education that fosters not only academic growth but also emotional and social development.

AI systems, despite their advanced capabilities, cannot replicate the empathy, encouragement, and nuanced understanding that human teachers provide. Studies argues that over-reliance on AI in education might lead to a transactional approach to learning, where students are treated as data points rather than individuals with unique needs and aspirations. This depersonalization can hinder the development of critical thinking, creativity, and other higher-order cognitive skills that thrive in a more human-centered educational environment (Anon 2024).

Furthermore, the implementation of AI can lead to standardized learning experiences that do not account for individual differences. While AI can adapt content based on performance metrics, it often lacks the ability to understand the broader context of a student's life, such as personal challenges or motivational factors, which can significantly impact learning. Thus, educators must strike a balance between leveraging AI for efficiency and maintaining the human touch that is essential for holistic education.

5.4 Accountability

Accountability refers to the responsibility of developers, educators, and institutions to ensure that AI systems are used ethically and transparently, and to address any potential harms that may arise from their use. It involves ensuring that AI systems contribute positively to the learning experience, do not perpetuate biases, and have clear channels for addressing grievances. Accountability is also described as the practice of ensuring that the creators of automated decision-making systems are held liable for the outcomes produced by their technology (Memarian & Doleck 2023).

5.5 Transparency

Transparency in AI systems is essential for fostering trust, understanding, and accountability. When AI systems are transparent, it helps users understand the rationale behind decisions, which is particularly important in educational settings where understanding the decision-making process can aid in learning. This clarity is achieved through clear documentation that explains the algorithms used, the data sets, and the decision-making process (Oneil, 2016).

While AI offers promising opportunities to enhance educational processes, it is imperative to address concerns related to data privacy, algorithmic bias, depersonalization of learning, transparency, and inequality. Ensuring robust data protection measures is essential for maintaining student privacy and trust. Actively working to eliminate biases in AI algorithms is crucial for promoting fairness and reducing inequality.

Additionally, maintaining the essential human elements of education is vital to prevent the depersonalization of learning and ensure meaningful student engagement. Promoting transparency in AI systems will help to build accountability and trust. By addressing these concerns, educators and policymakers can harness the benefits of AI while safeguarding the rights and needs of all students, fostering an equitable and effective educational environment.

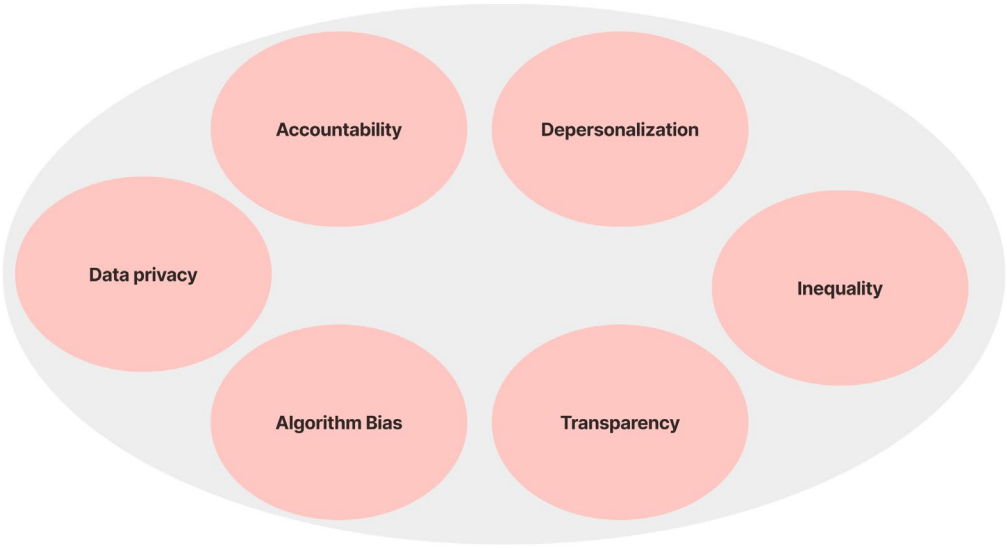


Figure 2: AI in Education: Ethics concerns

6. Conclusion

The integration of AI in education is transforming the landscape of learning and teaching. AI has introduced a new era of personalized learning, enabling tailored educational experiences that cater to the individual needs of each student. The evolution of intelligent tutoring systems, from their early days as rule-based programs to their current sophisticated architectures incorporating advanced algorithms and machine learning, highlights

the impact of AI on education. These systems are designed to provide real-time, adaptive support, making learning more effective.

GenAI further enhances the educational experience by creating dynamic content, simulating real-world scenarios, and offering interactive learning opportunities. This enriches the learning process and makes it more engaging and adaptable to various learning preferences. Additionally, AI can augment the role of educators by automating administrative tasks, providing detailed insights into student performance, and enabling more personalized teaching strategies. However, while the potential benefits are substantial, the integration of AI in education is fraught with numerous challenges.

Educators often encounter obstacles such as insufficient training, resistance to change, and the complexity of implementing new technologies. Overcoming these challenges requires ongoing professional development, robust support systems, and a collaborative approach to technology adoption. Without these, the effectiveness and acceptance of AI tools in the classroom may be limited.

Moreover, the deployment of AI in education raises several critical ethical considerations. Ensuring data privacy is paramount, given the vast amounts of student data collected and analyzed by AI systems. Ethical data-handling practices are crucial to maintaining trust and compliance with regulatory standards. Algorithmic bias is another significant concern, as AI systems can inadvertently reinforce existing biases present in their training data, leading to unfair or discriminatory outcomes. Developing strategies to identify, monitor, and mitigate these biases is essential to ensure equity and fairness in AI-driven education.

Furthermore, while AI enhances personalized learning, there is a risk that it may depersonalize the educational experience by reducing human interaction, which is essential for social and emotional learning. The balance between technology and human touch is delicate and must be carefully managed to preserve the holistic development of students.

In summary, AI's integration into education offers significant benefits, including personalized learning and enhanced support for educators. However, it is imperative to address the associated challenges and ethical considerations to maximize AI's potential in fostering educational excellence and innovation. By thoughtfully implementing AI and continuously evaluating its impact, educational institutions can create an environment where AI serves as a powerful tool for enhancing the learning experience while supporting the educational community. Nonetheless, this path requires vigilance, adaptation, and a commitment to addressing the complexities that AI introduces to the educational landscape.

References

Anderson, J. R. (1993). Problem solving and learning. The American Psychologist, 48(1), 35–44. https://doi.org/10.1037/0003-066x.48.1.35

Anon (2024) Using GenAI Effectively in Higher Education: Sustainable and Ethical Practices for Learning, Teaching and Assessment. 1st edition. [Online]. Routledge.

Baker, R. S. & Inventado, P. S. (2016) "Educational Data Mining and Learning Analytics: Potentials and Possibilities for Online Education, " in Emergence and Innovation in Digital Learning. Athabasca: Athabasca University Press. pp. 81–98.

Bloom, B. S. (1984) "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," Educational researcher (Washington, D.C.: 1972), Vol13, No.6, p. 4. doi: 10.2307/1175554.

Carlson, M. et al. (2024) "Utilizing OpenAI's GPT-4 for written feedback", TESOL journal, Vol 15, No.2. doi:10.1002/tesj.759

Carmon, C. M. et al. (2023) "Automated assessment of initial answers to questions in conversational intelligent tutoring systems: Are contextual embedding models really better?," Electronics, 12(17), p. 3654. doi: 10.3390/electronics12173654.

Cheng, Y. and Jiang, H. (2020). How do AI-driven chatbots impact user experience? Examining gratifications, perceived privacy risk, satisfaction, loyalty, and continued use. Journal of Broadcasting & Electronic Media, 64(4), pp.592-614.

Chow, J. C. L. et al. (2023) "Developing an AI-assisted educational chatbot for radiotherapy using the IBM Watson Assistant platform," Healthcare (Basel, Switzerland), Vol 11, No.17, doi: 10.3390/healthcare11172417.

Eynon, R. (2024) "Algorithmic bias and discrimination through digitalisation in education: A socio-technical view," in World Yearbook of Education 2024, pp. 245–260.

Graesser, A. C. (2017) "Assessment with computer agents that engage in conversational dialogues and trialogues with learners. Computers in human behavior," pp. 76607–76616.

Law, L. (2024) "Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review," Computers and Education Open, 6(100174), p. 100174. doi: 10.1016/j.caeo.2024.100174.

Memarian, B. and Doleck, T. (2023) "Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review," Computers and Education: Artificial Intelligence, 5(100152), p. 100152. doi: 10.1016/j.caeai.2023.100152.

OpenAI et al. (2023) “GPT-4 Technical Report.” doi: 10.48550/ARXIV.2303.08774.
Oneil, C. (2016) Weapons of math destruction how big data increases inequality and threatens democracy.
Raj, N. S., & Renumol, V. G. (2022) A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020. Journal of Computers in Education, 9(1), 113–148. https://doi.org/10.1007/s40692-021-00199-4
Vanlehn, K. (2011) “The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems,” Educational Psychologist, Vol 46, No. 4, pp. 197–221.

From Data to Decisions: Leveraging AI for Proactive Education Strategies

Willie Moore and Li-Shiang Tsay
North Carolina A&T University, Greensboro, USA

Wdmoore@aggies.ncat.edu
ltsay@ncat.edu

Abstract: The advancement of Artificial Intelligence (AI) and Large Language Models (LLMs) ushers in a new era in education, characterized by more adaptive, personalized learning experiences. This literature review examines the profound impact of these technologies on student engagement, achievement, and personalized learning within higher education institutions. Through a systematic analysis of scholarly articles from 2022 to 2024, this review explores how AI is reshaping educational practices through enhanced feedback mechanisms, predictive analytics, and innovative teaching methodologies. The findings indicate that AI significantly improves student support services by enabling early identification of at-risk students and by facilitating tailored educational interventions. Moreover, the deployment of chatbots and LLMs, such as GPT (generative pre-trained transformer) and BERT (bidirectional encoder representations from transformers), offers promising enhancements in instructional strategies and student assessments, fostering richer, interactive learning environments. However, the integration of these technologies also introduces ethical challenges, necessitating consideration of issues such as data privacy and bias. The review emphasizes the need for ethical frameworks and responsible AI usage to ensure technology enhances educational outcomes without compromising fairness or integrity. Future research directions are suggested, focusing on broader AI applications across various educational settings and the need for longitudinal studies to assess the long-term effects of AI integration in education.

Keywords: Educational AI, LLMs, Personalized learning, Predictive analytics, AI feedback, Chatbots

1. Introduction

AI innovations shape our daily lives, from Alexa to personalized education. The release of ChatGPT in late 2022 redefined AI, demonstrating machines' ability to learn human language and perform tasks like text generation, summarization, question answering, text classification, translation, and prediction. In education, AI and LLMs mark a significant evolution, introducing adaptive and customized learning experiences, automated grading, feedback systems, and intelligent tutoring (Wangwiwattana and Tongvivat, 2023). These technologies address student engagement, persistence, and achievement challenges but also raise ethical concerns. This study explores AI's impact on academic success, focusing on two research questions:

RQ1: How can institutions use AI to enhance student engagement and learning outcomes

RQ2: What are the implications of AI for personalized learning in higher education.

This systematic review analyses the current state and future potential of AI in education, using 172 articles from IEEE and ScienceDirect databases. The trend toward AI adoption in education aims to shape tomorrow's workforce. Our analysis shows AI technologies transforming the education environment through LLM integration, student success prediction, study preparation, and teaching. Higher education institutions use AI to personalize education, grade student work, and enhance language lessons. More research is needed to explore AI's broader impacts across educational levels and disciplines.

To guide readers through the findings and insights of this study, it is structured as follows: Section 2 discusses the methodology for data collection and screening. Section 3 examines the role of Large Language Models in enhancing educational practices, focusing on AI-Enhanced Feedback, Chatbot-Enhanced Learning, and Assessment. Section 4 delves into the application of predictive analytics to identify at-risk students. Section 5 focuses on the impact of AI tools on study preparation, using the FE exam as an example. Section 6 reviews AI applications in teaching, including enhancing motivation, classroom discourse, cognitive skills, and online engagement. Finally, Section 7 provides a summary of key findings, practical implications, and future research directions.

2. Systematic Review Approach

2.1 Data Collection

This literature review curated scholarly articles on Artificial Intelligence, Machine Learning, Large Language Models, and chatbots' educational outcomes, focusing on higher education. A systematic search and screening identified relevant literature from 2022 to 2024.

2.2 Search Strategy

An extensive search was conducted across two major databases, ScienceDirect and IEEE, using a set of predetermined search strings aimed at identifying a wide array of research focusing on the application of AI/ML/LLMs and technology adoption in education, especially within the context of minority education. The search strings were as follows:

- "machine learning" AND Education AND minority
- ("artificial intelligence in higher education") OR (HBCU AND Technology) OR (“technology adoption” AND minority) OR ("machine learning" AND Education AND minority)
- “large language model” AND education

This search yielded a total of 110 articles from IEEE and 186 from ScienceDirect. After merging the two lists and removing duplicates, 245 articles were identified. A further filter was applied to select articles published within the years 2022 to 2024, resulting in a refined list of 172 articles.

2.3 Screening Process

Preliminary screening ensured article’s relevance and quality based on four key areas:

1. AI Relevance: Does the article provide significant insights into AI technologies that are applicable to educational settings, even if not directly applied within them?
2. Focus on Higher Education: Is the primary setting of the study or discussion within higher education institutions?
3. Empirical Evidence: Conceptual papers, while valuable for theoretical discussions, often lack the empirical data necessary to evaluate AI’s measurable impact on student outcomes. Does the article provide empirical evidence or thorough analysis of technology/AI's impact on student outcomes?
4. Relevance to Minority Education: Does the article directly reference Historically Black Colleges and Universities or other minority-serving institutions?

Each response of “Yes” to any of the screening questions was scored as 1 point. Articles scoring at least 3 points in total were selected for initial inclusion. This process yielded 56 articles

2.4 Article Grouping

After preliminary screening, articles were reviewed for relevance to AI/ML/LLMs/chatbots in improving educational outcomes. They were categorized by theme, resulting in 26 articles for the review.

2.5 Important Notes

The original design of this literature review was to focus on the application of AI technologies in Historically Black Colleges and Universities. During the literature search, it became clear that there was a significant gap in research directly addressing the use of AI within HBCUs. The limited availability of empirical studies in this context necessitated expanding the scope of the review to include general higher education settings.

This review delves into four critical areas where AI technologies, including Machine Learning and Large Language Models, contribute to educational strategies: LLM Educational Integration, Prediction – Student Success, Study Preparation, and Teaching. Each section explores different dimensions of AI application.

3. LLM Educational Integration

Large Language Models (LLMs), such as GPT and BERT, have transformed educational practices by enhancing teaching and learning processes. These AI systems, trained on extensive datasets, are adept at understanding and generating human-like text, making them invaluable in educational settings (Dai et al., 2023; Essel et al., 2023; Niloy et al., 2024; Bewersdorff et al., 2023). This section explores LLM integration in education, focusing on AI-Enhanced Feedback, Chatbot-Enhanced Learning, and Assessment.

LLMs support educators by automating student error analysis, identifying learning gaps, and tailoring educational responses. Bewersdorff et al. (2023) demonstrated that LLMs help create individualized learning strategies addressing each student's unique needs. These tools improve efficiency while allowing educators autonomy in their use of AI systems, enhancing teaching without undermining professional discretion (Bewersdorff et al., 2023; Ouatiq et al., 2022).

Ethical implications are critical in integrating LLMs in education. AI tools must enhance educational practices while addressing ethical concerns (Vetter et al., 2024). The introduction of ChatGPT in academic settings has spurred ethical discussions and empirical studies on its use by students (Niloy et al., 2024). This underscores the need to balance benefits and challenges, ensuring AI implementation is effective and responsible.

3.1 AI Enhanced Feedback

Providing effective feedback is a critical component of enhancing student engagement and learning outcomes. The role of AI-Enhanced Feedback becomes particularly relevant in addressing how institutions can utilize AI to improve these aspects. The emergence of LLMs has resulted in the creation of automated, evidence-based feedback systems, presenting a promising solution for generating such feedback (Fleckenstein et al., 2023; Meyer et al., 2023; Dai et al., 2023).

Such feedback has been shown to significantly improve students' revision performance and task motivation. As noted by Myer, “Students who received feedback generated by the LLM improved to a greater degree in the text revision compared to students in the control group” (Meyer et al., 2023, p. 6). This demonstrates LLMs' capacity to enhance student learning experiences and outcomes effectively.

In educational technology, feedback plays a crucial role in improving student learning outcomes, as it is widely acknowledged as an effective technique for improving student learning (Dai et al., 2023). This foundational role of feedback underlines its importance in the learning process, where the integration of AI tools supplemented with large language models like ChatGPT represents a significant advancement over traditional methods. For instance, in the context of computer programming, the deployment of AI as a virtual peer through ChatGPT was used to offer feedback in a study performed by Dos Santos and Cury (2023). Dos Santos and Cury (2023) discovered that using AI, such as ChatGPT, as a virtual peer within peer instruction models significantly boosted students' programming abilities and conceptual comprehension compared to conventional teaching methods.

This integration demonstrates the potential of AI to not only support basic learning tasks but also to deepen student engagement with the subject matter, directly addressing the question of how AI can be utilized to enhance educational outcomes. The application of AI-Enhanced feedback directly addresses research question one by demonstrating how AI can improve student engagement and learning outcomes through personalized feedback mechanisms.

3.2 Chatbot Enhanced Learning

Enhancing student engagement and learning outcomes is crucial for educational success. AI-powered chatbots represent a significant evolution in educational environments by supporting active and self-regulated learning (SRL) strategies. Advanced tools like generative AI chatbots, such as ChatGPT and Bing Chat, provide personalized support, feedback, and guidance, fostering deeper learning (Essel et al., 2023; Lin and Chang, 2023; Meyer et al., 2023). Chatbots can simulate roles such as tutor, teaching assistant, and peer, offering round-the-clock assistance and enhancing the educational process (Yildirim-Erbasli and Bulut, 2023). It’s been stated “By leveraging the potential of chatbot technology and its ability to provide personalized support, feedback, and guidance, educators can create more engaging and effective learning environments that promote active learning and SRL” (Lin and Chang, 2023, p. 6).

The ethical integration of such technology involves ensuring responsible use, enhancing educational outcomes without compromising human interactions or leading to over-reliance on technology. This requires ongoing assessment and adaptation to align with ethical standards, safeguarding educational integrity and student welfare.

Chatbot-enhanced learning addresses both research questions by demonstrating how AI can enhance student engagement and learning outcomes (research question 1) and by exploring the implications of AI for personalized learning in higher education (research question 2). This section highlights the potential of chatbots to improve student engagement and offer learning experiences on a individualized level.

3.3 Assessment

As we explore the integration of Large Language Models within the educational framework, the realm of assessment reveals both transformative potential and challenges that need careful navigation. This aligns directly with the themes of improving educational outcomes and addressing the ethical implications of AI for personalized learning, as stipulated in both research questions. In the field of educational assessments, the complexity of evaluating student work, such as experimentation protocols, is amplified by the diverse nature of

the data. Bewersdorff et al. (2023) described the difficulty of detecting "logical errors in complex, incomplete, contradictory, and heterogeneous data, such as students' experimentation protocols". This underscores the necessity for advanced tools capable of handling such complexities. LLMs, through their sophisticated text analysis capabilities, provide a solution that could potentially streamline the assessment process. AI's role in education extends beyond mere auto-grading, offering detailed insights and feedback that can significantly enhance the quality of student assessments and learning experiences (Yildirim-Erbasli and Bulut, 2023).

3.3.1 Writing evaluation

LLMs facilitate the creation of more reflective and engaging learning environments. Their ability to generate high-quality reflective writings supports students in engaging more deeply with learning materials. Li et al., (2023) argue that advanced generative LLMs and related tools have the potential to produce high-quality reflective writing examples. This application of LLMs illustrates their versatility in fostering higher-level thinking and self-reflection, which are critical components of successful learning strategies.

In writing evaluation, LLMs are increasingly employed to provide detailed feedback, which directly addresses our second research question. This feedback can significantly enhance the learning process. The deployment of LLMs in assessments goes beyond simply automating grading. It also aims to enhance students' assessment experiences by using conversational agents to administer tasks and provide feedback (Yildirim-Erbasli and Bulut, 2023). These agents, designed for educational purposes, simulate human-like interactions, making the assessment process more engaging and less daunting for students.

3.3.2 Ethical use in assessment

The ethical use of AI in assessments raises several important concerns, particularly around fairness, academic integrity, and the authenticity of student work. The capabilities of LLMs pose significant challenges to maintaining the integrity of student submissions (Vetter et al., 2024). Human instructors cannot consistently differentiate between AI-generated text and student-authored content, further complicating the issue. Current AI detectors, while improving, are far from flawless, with notable misclassification rates that either falsely flag authentic student work or fail to identify AI-generated content (Fleckenstein et al., 2023). These limitations highlight the inadequacy of relying solely on detection tools to address the problem of academic dishonesty.

Rather than attempting to outlaw the use of AI in education, educators should consider the development of a comprehensive framework that integrates AI tools into their teaching practices (Vetter et al., 2024). Such a framework would allow instructors to adapt their pedagogy, using AI to enhance learning objectives rather than viewing it as a threat to academic integrity. This pedagogical shift should include clear guidelines for the ethical use of AI, ensuring that students understand the benefits and boundaries of these tools in the learning environment (Fleckenstein et al., 2023; Vetter et al., 2024).

The ethical implications of AI use in education cannot be ignored. AI tools, while powerful, raise questions about dependency on technology and the authenticity of student work (Vetter et al., 2024). Ensuring that AI supports educational goals without compromising assessment quality or fairness is critical. Educators must establish clear guidelines and ensure transparency in the responsible use of AI, balancing the benefits of these technologies with the need to uphold academic integrity, motivate students and foster genuine learning while preventing misuse (Yildirim-Erbasli and Bulut, 2023).

Writing Evaluation and Ethical AI Use in Assessment sections show that while LLMs can enhance assessment efficiency and effectiveness, their integration must focus on learning outcomes and ethical standards (Vetter et al., 2024). This alignment with academic objectives and ethical considerations addresses research question two.

4. Prediction – Student Success

It is essential to understand the factors that influence student success to improve educational outcomes. Predictive analytics in education aims to enhance student outcomes and institutional effectiveness by identifying risk factors for underperformance or dropout, addressing our second research question. These analytics are used across various educational levels and regions, demonstrating their broad appeal and necessity (Alwarthan, Aslam and Khan, 2022; Qin et al., 2023). Beyond immediate academic outcomes, predictive analytics impact long-term educational and career trajectories, as dropout negatively affects job opportunities and long-term earning potential (Manir and Deshpande, 2023). Addressing dropout factors can secure better futures for students. Research across universities highlights predictive analytics' transformative potential. Studies at

various institutions, from large public universities to small private colleges, aim to understand and mitigate student attrition factors.

Minority student underachievement in higher education remains significant and troubling, requiring urgent attention (Manir and Deshpande, 2023). This emphasizes the need for targeted interventions facilitated by predictive analytics. A study at Marquette University, which is not an HBCU, demonstrated effective use of machine learning algorithms to monitor and predict academic success, particularly focusing on demographic factors that contribute to the academic success of minority students. This study highlights the relationship between academic achievement and demographic factors, providing valuable insights into how predictive analytics can support minority students (Manir and Deshpande, 2023).

Marquette University's study also showed the efficacy of sophisticated machine learning models like Random Forests in predicting student outcomes. Random Forest achieved the highest accuracy for undergraduate and graduate datasets, with 96.7% and 91.1%, respectively (Manir and Deshpande, 2023). These models handle complex datasets and uncover hidden patterns, enhancing performance (Manir and Deshpande, 2023).

Other institutions have used various predictive models with differing success, often depending on data quality and educational context. Deep learning methods to analyse data from learning management systems (LMS) provide insights to refine academic programs and support services (Souai et al., 2022; Sagala et al., 2022; Vaarma and Li, 2024; Pecuchová and Drlík, 2023; Vives et al., 2024; Hlosta et al., 2022). The importance of Learning Management Systems (LMS) in predictive analytics is increasingly recognized. LMS data provides insights into student engagement and progress. Vaarma and Li (2024) suggest that LMS data features have strong predictive power and should be evaluated as independent variables in dropout predictions, facilitating better-targeted interventions (Vaarma and Li, 2024).

Predictive models to identify students at risk of academic failure enhance student performance in higher education (Pecuchová and Drlík, 2023). These models play a crucial role in early intervention strategies, significantly enhancing student support mechanisms (Souai et al., 2022; Sagala et al., 2022; Vaarma and Li, 2024; Pecuchová and Drlík, 2023; Vives et al., 2024; Hlosta et al., 2022).

However, traditional models like Random Forest and logistic regression have limitations. Individual classification algorithms have performance limits, which can be overcome by using ensembles of classifiers (Pecuchová and Drlík, 2023). This indicates a shift toward ensemble methods that integrate multiple models to enhance prediction accuracy and reliability.

Predictive models must be carefully designed to consider data availability and relevance. Even with limited data, carefully selected indicators can yield reliable predictions (Souai et al., 2022; Sagala et al., 2022; Vaarma and Li, 2024; Pecuchová and Drlík, 2023; Vives et al., 2024; Hlosta et al., 2022). Pecuchová and Drlík (2023) demonstrated that even with a small sample size, carefully selected indicators can assess various performance metrics, making predictive analytics accessible to smaller educational settings.

Challenges arise from the unpredictable nature of student life, affecting learning outcomes and retention. Hlosta et al. (2022) found that unexpected events during studies significantly impact student behaviour. Recognizing these variables is crucial for developing adaptive and responsive predictive models.

Advancements in deep learning technologies offer new avenues for predictive analytics in education. A study on tertiary education found that deep learning methodologies can analyse student performance. Researchers proposed expanding the study by incorporating additional interpretable factors (Hasib et al., 2022). This highlights the potential for integrating more complex algorithms to delve deeper into predictive factors, suggesting future research directions.

5. Study Preparation

Enhancing study preparation is essential for academic success, particularly in preparing for critical examinations. Advancements in AI have shown significant promise in transforming various aspects of education, including study preparation. This area addresses both research questions by showcasing how AI tools can enhance study preparation, improve learning outcomes, and provide personalized study experiences.

In a recent study that investigated the potential of AI tools like ChatGPT for preparing students for critical examinations, several notable findings were revealed. This study specifically focused on the US Fundamentals of Engineering (FE) exam, a key steppingstone for engineering students in the United States. The purpose of the study by Pursnani et al. (2023) is to provide insights into the effective use of AI in preparing for standardized

exams. By leveraging AI tools, educators can offer more personalized and adaptive learning experiences that cater to the specific needs of students (Pursnani et al., 2023).

This finding suggests that while AI may not yet fully replace human tutors, it serves as a valuable complement, particularly in areas where scalable and individualized feedback is beneficial. This study's value rests in its ability to be applied across various academic disciplines, not just engineering. These findings could have broader implications, potentially benefiting a wide range of educational fields by adapting AI tools for different types of exam preparations. Pursnani et al. (2023) observed both the significant potential and limitations of AI in education, indicating that its role in this field is continually advancing. They acknowledge that while AI can significantly enhance study preparation, ongoing refinement and customization are necessary to maximize its effectiveness and applicability in diverse educational settings.

6. Teaching

This section addresses the first research question: How can institutions use AI to enhance student engagement and learning outcomes? By exploring various aspects of AI applications in teaching, we examine how these technologies can foster motivation, improve classroom interactions, and develop cognitive skills.

6.1 AI Enhanced Motivation

Motivation is a critical factor in educational success. Lai, Cheung, and Chan (2023) explored how AI tools can influence student motivation. Their study found that AI technologies can significantly enhance motivation, leading to greater engagement and more favourable attitudes towards learning. This underscores the potential of AI to boost student engagement by fostering motivation, directly addressing the first research question.

6.2 Classroom Discourse

Improving classroom interactions is another vital aspect of enhancing student engagement. Wang, Yang, and Chen (2023) conducted a systematic review on the impact of AI in classroom discourse. Their findings indicate that AI tools consistently enhance classroom dynamics, leading to greater engagement and improved student behaviours. This demonstrates how AI can create more interactive and lively learning environments, further addressing the first research question.

6.3 Cognitive Skills Development via ChatGPT

Research on the cognitive skills of undergraduate students using ChatGPT showed significant improvements. Essel et al. (2023) found that students who used ChatGPT demonstrated substantial enhancement in critical thinking skills compared to those who did not. This highlights the role of AI tools in fostering cognitive development, contributing to better learning outcomes and addressing the first research question.

6.4 Predictive Models for Online Learning

In online education, understanding and predicting student engagement is crucial. Verma et al. (2023) demonstrated the effectiveness of AI models in analysing teacher behaviours and movements in video conferencing to gauge student engagement. This use of AI to tailor educational content dynamically supports and enhances student engagement in online learning environments, aligning with the first research question.

7. Key Insights

This literature review highlights several significant impacts of AI and Large Language Models in education, which provide a comprehensive understanding of how these technologies can transform educational practices. The key insights from the review are as follows:

- Enhanced Personalized Learning: AI and LLMs create adaptive and personalized learning environments. These technologies allow for tailored educational experiences, improving student engagement and learning outcomes by addressing individual needs.
- Predictive Analytics for Student Success: The use of machine learning and predictive analytics effectively identifies at-risk students, enabling early interventions. These tools help institutions to proactively support students, reducing dropout rates and improving overall academic success.
- Innovative Teaching Methodologies: AI tools such as chatbots and virtual peers enhance teaching methodologies. By providing instant feedback, personalized support, and fostering interactive learning environments, these tools improve student motivation and cognitive skills.

- Improved Assessment Practices: LLMs streamline and enhance assessment processes by automating the evaluation of student work and providing detailed, evidence-based feedback. This improves the efficiency and effectiveness of assessments, supporting better learning outcomes.
- Ethical Considerations: The integration of AI in education raises important ethical issues, including data privacy, bias, and the authenticity of student work. There is a critical need for ethical frameworks to ensure responsible AI usage that maintains fairness and integrity in educational practices.
- Broader Implications for Minority Education: There is a noted gap in the literature concerning the application of AI in Historically Black Colleges and Universities.

8. Conclusion

The incorporation of Artificial Intelligence and Large Language Models in education is reshaping traditional educational practices, leading to more adaptive and personalized learning experiences. This literature review has highlighted several key areas where AI technologies have significantly impacted student engagement, achievement, and personalized learning within higher education institutions. By providing enhanced feedback mechanisms, predictive analytics, and innovative teaching methodologies, AI is transforming the educational landscape.

8.1 Implications for Practice

The results of this review offer several practical implications for educators and institutions:

- Integrating AI Tools: Educators should leverage AI to enhance interactive learning and support student engagement. These tools can provide immediate, personalized feedback, helping to keep students motivated and involved in the learning process.
- Customized Learning Plans: AI can offer personalized feedback and assessment, allowing educators to tailor educational content to meet individual student needs. This customization can help address diverse learning styles and improve overall student outcomes.
- Ethical Deployment: Institutions must address the ethical challenges associated with AI integration, such as data privacy, bias, and the authenticity of student work. Establishing clear ethical guidelines and frameworks is crucial to ensure that AI enhances educational outcomes without compromising integrity or fairness.

8.2 Future Research Directions

Future research should explore several areas to fully comprehend and employ the capabilities of AI in education:

- Broad AI Applications: More research is needed on AI applications across different educational levels and disciplines to understand the broader impacts. Investigating how AI can be used in primary, secondary, and higher education will provide insights into its versatility and effectiveness.
- Longitudinal Studies: Long-term studies are necessary to assess the effects of AI integration on student retention and success.
- HBCU Education: More focused research on AI applications in Historically Black Colleges and Universities is needed. Understanding how AI can support these institutions and address their unique challenges will be crucial for ensuring equitable access to AI-enhanced educational opportunities.

8.3 Final Thoughts

In conclusion, AI and LLMs hold great potential to transform education by fostering more personalized and engaging learning environments. However, it is critical to tackle the ethical considerations and ensure responsible AI usage to maximize the benefits for all students. By continuing to explore and understand the implications of AI in education, institutions can effectively leverage these technologies to enhance educational outcomes and student success.

References

Alwarthan, S., Aslam, N. & Khan, I.U. 2022, An Explainable Model for Identifying At-Risk Student at Higher Education, Institute of Electrical and Electronics Engineers (IEEE).

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E. & Nerdel, C. 2023, Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters, Elsevier BV.

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y., Gašević, D. & Chen, G. 2023, Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT, IEEE.

Dos Santos, O.L. & Cury, D. 2023, Challenging the Confirmation Bias: Using ChatGPT as a Virtual Peer for Peer Instruction in Computer Programming Education, IEEE.
Essel, H.B., Vlachopoulos, D., Essuman, A.B. & Amankwa, J.O. 2023, ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs), Elsevier BV.
Fleckenstein, J., Meyer, J., Jansen, T., Keller, S.D., Köller, O. & Möller, J. 2024, Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays, Elsevier BV.
Hasib, K.M., Rahman, F., Hasnat, R. & Alam, M.G.R. 2022, A Machine Learning and Explainable AI Approach for Predicting Secondary School Student Performance, IEEE.
Hlosta, M., Herodotou, C., Papathoma, T., Gillespie, A. & Bergamin, P. 2022, Predictive learning analytics in online education: A deeper understanding through explaining algorithmic errors, Elsevier BV.
Lai, C.Y., Cheung, K.Y. & Chan, C.S. 2023, Exploring the role of intrinsic motivation in ChatGPT adoption to support active learning: An extension of the technology acceptance model, Elsevier BV.
Lin, M.P. & Chang, D. 2023, CHAT-ACTS: A pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning, Elsevier BV.
Manir, S.B. & Deshpande, P. 2023, Student success analysis for minority students in higher education, IEEE.
Meyer, J., Jansen, T., Schiller, R., Liebenow, L.W., Steinbach, M., Horbach, A. & Fleckenstein, J. 2023, Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions, Elsevier BV.
Niloy, A.C., Bari, M.A., Sultana, J., Chowdhury, R., Raisa, F.M., Islam, A., Mahmud, S., Jahan, I., Sarkar, M., Akter, S., Nishat, N., Afroz, M., Sen, A., Islam, T., Tareq, M.H. & Hossen, M.A. 2024, Why do students use ChatGPT? Answering through a triangulation approach, Elsevier BV.
Ouatiq, A., Faris, S., Mansouri, K. & Qbadou, M. 2022, Towards the Use of teacher-chosen features to predict students' performance, IEEE.
Pecuchova, J. & Drlik, M. 2023, Predicting Students at Risk of Early Dropping Out from Course Using Ensemble Classification Methods, Elsevier BV.
Pursnani, V., Sermet, Y., Kurt, M. & Demir, I. 2023, Performance of ChatGPT on the US fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice, Elsevier BV.
Qin, K., Xie, X., He, Q. & Deng, G. 2023, Early Warning of Student Performance With Integration of Subjective and Objective Elements, Institute of Electrical and Electronics Engineers (IEEE).
Sagala, T.M.N., Permai, S.D., Gunawan, A.A.S., Barus, R.O. & Meriko, C. 2022, Predicting Computer Science Student's Performance using Logistic Regression, IEEE.
Souai, W., Mihoub, A., Tarhouni, M., Zidi, S., Krichen, M. & Mahfoudhi, S. 2022, Predicting At-Risk Students Using the Deep Learning BLSTM Approach, IEEE.
Vaarma, M. & Li, H. 2024, Predicting student dropouts with machine learning: An empirical study in Finnish higher education, Elsevier BV.
Verma, N., Getenet, D.S., Dann, D.C. & Shaik, T. 2023, Designing an artificial intelligence tool to understand student engagement based on teacher's behaviours and movements in video conferencing, Elsevier BV.
Vetter, M.A., Lucia, B., Jiang, J. & Othman, M. 2024, Towards a framework for local interrogation of AI ethics: A case study on text generators, academic integrity, and composing with ChatGPT, Elsevier BV.
Vives, L., Cabezas, I., Vives, J.C., Reyes, N.G., Aquino, J., Cóndor, J.B. & Altamirano, S.F.S. 2024, Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks, Institute of Electrical and Electronics Engineers (IEEE).
Wang, D., Tao, Y. & Chen, G. 2023, Artificial intelligence in classroom discourse: A systematic review of the past decade, Elsevier BV.
Wangwiwattana, C. & Tongvivat, Y. 2023, Automating Academic Assessment: A Large Language Model Approach, IEEE.
Yildirim-Erbasli, S.N. & Bulut, O. 2023, Conversation-based assessment: A novel approach to boosting test-taking effort in digital formative assessment, Elsevier BV.

Generative AI and its Impact on Activities and Assessment in Higher Education: Some Recommendations From Master's Students

Peter Mozelius
Mid Sweden University, Sundsvall, Sweden

Peter.Mozelius@miun.se

Abstract: The rapid development of generative AI (GenAI) raises new questions in higher education such as: What should be the university policy regarding GenAI? How ought courses be redesigned for fair and resilient assessment? What the added pedagogical and didactical values when involving GenAI in teaching and learning activities? Different universities have rapidly created and presented contradictory standpoints and draft policies, and teachers show different opinions regarding the pros and cons of GenAI. This study has been carried out with a student perspective, where 16 students have been examining their own Master's programme on sustainable information provision. Students have assessed the assessment in their previous courses in the Master's programme. The aim of the study is to investigate how sustainable course activities and assignment are, and to explore how GenAI tools might support and facilitate teaching and learning activities. Moreover, the students were given the task to test detection software on GenAI generated solutions to assignments in chosen Master's courses. Students conducted these tasks as a part of a 7.5 ECTS project course in the same Master's programme as the investigated courses are a part of. For inspiration and for background information on artificial intelligence to the project work students participated in the first Symposium on AI Opportunities and Challenges (SAIOC) in December 2023. Data have been gathered from reports of 3 group projects where 16 students have investigated 5 freely chosen courses in the programme in each group work. Beside from testing GenAI tools in existing activities and assignments students also interviewed the subject matter experts that are responsible for the chosen courses. Results were firstly analysed and presented in group reports, combined with 16 individual reflection essays. Regarding the individual essays students were instructed to bring up ethical perspectives on GenAI in higher education, and also to present and discuss suggestions for how the current course design and assignments better could be redesigned for improved sustainability and fairness. Finally, all the group reports and the individual reflection essays were thematically analysed by the author, who also is the subject matter expert and main teacher for the project course. Findings show that many of the existing assignments in the Master's programme could be partly solved with different GenAI tools. The AI generated solutions showed different levels of quality and correctness for different types of activities and assignments. An ethical concern that many student essays brought up was the relatively poor quality of the tested detection software. A question in one of the essays was if teachers should use detection software with an accuracy rate just above 50% to evaluate student submissions. The recommendations from both the students and the author are to provide clear instructions about when GenAI is allowed and not in course activities, and to redesign the course structure for continuous assessment. With or without GenAI tools, a continuous assessment where the whole study path through a course is assessed, and not only isolated submissions, would strengthen fairness and sustainability. Finally, several students suggest oral examinations as a complement to the existing assessment methods, even if their findings showed that GenAI tools can be used to prepare oral presentations.

Keywords: Artificial intelligence, Generative AI, GenAI, Higher education, Sustainable education

1. Introduction

During the last years there have been a rapid development of generative AI (GenAI), with new GenAI tools that raises new questions in higher education. Research studies have reported on the need of new assessment methods to secure that students submissions are original work (Luo, 2024). The ongoing third spring of AI might be a hype followed by a new AI winter, but the impact on teaching and learning in higher education will remain (Pons, 2023; Whitham et al., 2023). Some universities have rapidly created draft policies contradicting each other, while other universities have kept a passive silence. As discussed in Luo (2024), to strict and banning policies may result in that students become hesitant to use GenAI even for legitimate learning purposes.

Higher education and technological development have a symbiotic relationship that goes way back. On one hand, technological development could be seen as a result of research efforts that has led to technological applications in society. One such field is digital technologies for technology enhanced learning which currently has a strong focus on GenAI technologies. On the other hand, these technologies are also applied in experiments and projects that aim to further develop higher education operations and to investigate how AI impacts learning and teaching in higher education. There is a strong need today for higher education initiatives on how to involve GenAI in course design and students' learning activities. (Mozelius et al., 2024)

Several research studies have investigated the teacher perspective of GenAI in educational contexts (Duc 2023; Stewart, 2023), and pointed out the need for teacher professional development on the use of GenAI in education (Mishra, Warr & Islam, 2023; Chiu, 2024). As highlighted by Chan and Hu (2023) it is important to bring in the student voices in the discussion as well. In this study 16 students have examined their own Master's programme

on sustainable information provision, but that the teacher voices have been involved as well when students have conducted semi-structured interviews with the teachers have designed the ingoing courses. The aim of the study is to investigate the sustainability of course activities and assignment in the Master The aim of the study is to investigate how sustainable course activities and assignment are, and to explore how GenAI tools might support and facilitate teaching and learning activities in the Master's programme.

2. Method

The overall research approach could be described as educational action research as outlined by Norton (2009). A branch of action research in the British tradition "*that links research to improvement of practice and is education orientated*" Norton (2009, p. 71). Aligned to the idea that research results should have a positive impact on the redesign of the future versions of the investigated courses. As posit by Arnold and Norton (2021), an educational action research that should strive to bring change, innovation and educational development. The educational action research has been combined with the idea of participatory where students should not only be researched objects, but also partners in the research process (Baum, MacDougall and Smith, 2006). Moreover, the study had an overall qualitative approach were interviews and activity evaluations in the student groups have been summarised in group reports and individual assignments.

2.1 Research Context

This study was carried out with a student perspective, where students have examined the Master's programme on sustainable information provision that they are a part of. The overall purpose of the Master's programme is to provide advanced training in sustainable information services with special focus on management and design. For this study, the main idea has been that students should investigate the assessment and teaching and learning activities in their previous courses in the Master's programme. Students carried out this work as a part of a 7.5 ECTS project course where they also should plan and conduct a project in smaller groups with 4-6 participants. As inspiration and for background information on GenAI for the project work students participated in the first Symposium on AI Opportunities and Challenges, SAIOC (Jaldemark, Mozelius & Humble, 2023).

Besides from the students' findings, this study also involves the teacher perspective since all groups have interviewed at least one subject matter expert or course designer in one of the chosen Master's courses. Students were also given the task to test different detection software to compare GenAI generated solutions to human generated solutions to assignments in the chosen Master's courses. All interviews were conducted in Swedish and quotes have been translated from Swedish to English by the author.

2.2 Data Collection

Data have been collected in group projects where 16 students, divided into 3 groups, investigated 5 courses in each group work. All groups carried out teacher interviews, document studies and various GenAI tool testing. As a part of the tool testing student generated solutions to assignments were compared to AI generated solutions with the use of detection software. Students could freely choose detector tools from the listing in Figure 1 below. This is the actual listing in December 2023 during the course span, and that the current listing in July 2024 involves 10 detector tools with a slightly higher accuracy rate than the percentages in Figure 1.

Winston AI	84%	✓	4.2/5
Originality.AI	76%	✓	3.8/5
GLTR	72%	✓	3.6/5
Sapling	68%	✓	3.4/5
Content at Scale	66%	✓	3.3/5
Copyleaks	66%	✓	3.3/5
Crossplag	58%	✓	2.9/5

Figure 1: The 7 Greatest AI Content Detector Tools (Forbes, 2023, December)

In the individual essays students were also asked to bring up ethical perspectives on GenAI in higher education. Moreover, students should present suggestions on how the current course design and assessment better should

be redesigned for sustainability and fairness. Finally, the author has used notes taken at discussion webinars in the described project course.

2.3 Data Analysis

Firstly, results were analysed by students individually, and in groups, before the presentations at course webinars, in written group reports, and in individual reflection essays. The different groups have used different analysis methods, but the overall approach could be described as thematic analyses.

Secondly, the group reports and the individual essays were analysed by the author in an inductive thematic analysis. A process that followed the six-step process outlined by Braun and Clarke (2006) to identify themes, i.e. patterns in the data that can contribute to meeting the aim of the study. The six-steps that guided the analysis were: 1) To become familiar with the data, which started early during the discussions at the course webinars, 2) To generate initial codes, which also started out early with the student analyses, 3) To search for themes, which has been done both by the students and the author, 4) To review themes, which was done in a first iteration by the students, and then in a second iteration when student submissions were aggregated by the author, 5) Defining the final themes, was carried out by the author in an aggregation and reconstruction of the student themes, and finally 6) Writing up and creating a narration to present the findings was done by the author, but based on the student submissions.

2.4 Ethical Considerations

All involved stakeholders have been kept as anonymous as possible, and the ambition has been to not criticise individuals or specific courses. The overall principal has been to leave out all names and more specific details to achieve the aim of investigating how sustainable course activities and assignment are, and to explore how GenAI tools might support and facilitate teaching and learning activities.

3. Findings and Discussions

The findings from the analysis have been grouped into the categories of 'Activity types', 'Assignment complexity', 'Detection software', 'Ethical issues', and finally 'Recommendations for redesign'. The found themes are reported and discussed one by one here below.

3.1 Activity Types

In a programme with a wide variety of courses and course activities the findings were that most of the activities could be solved, or partly solved with the assistance of GenAI tools. One of the groups that used a wide variety of GenAI tools brought up that also the oral activities such as seminars and discussion activities could be prepared with AI support. The students must of course be the ones that presents and discusses, but that arguments and problematisation could be outlined in chatbot conversations. A conclusion in all group reports is that oral activities and activities where students are supposed to list basic facts or to explain concepts the GenAI support works well. This could be aligned to Bloom's taxonomy that is depicted in Figure 2, and that the lower levels of remembering and understanding are handled without errors by most of the tested tools. Furthermore, to draw connections on the 'Analyze level' of the taxonomy often show relevant output.

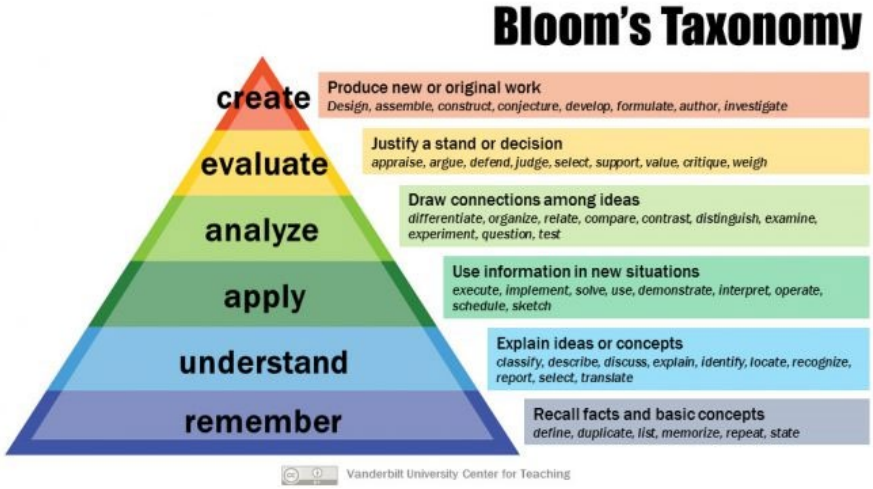


Figure 2: Bloom's taxonomy (Armstrong, 2010)

For other levels such as apply and create, one group report brings up that "The completely AI generated text often show poor quality, with minor errors, superficial analysis and incorrect or missing references". This aligns with the answers from the interviewed subject matter experts, where one of the interviewees brought up that "Two cases where students have used AI tools to solve assignments were mainly detected by the complete lack of references".

In the interview same the informant also pointed out that the use of AI tools for language correction is something to recommend, and something to place at the same level as using the spelling support in a word processing software. The group that had the largest toolbox also tested the free version of 'Grammarly', but their evaluation was that "The language support was not of higher quality than in a word processing software such as Microsoft Word". Another group, where several members have other mother tongues than Swedish, the group report highlighted the benefits of using different software to improve both the written Swedish and the written English in course submissions. Finally, two of the groups found DeepL to be a useful tool for translations between different languages. However, with a slight warning for "losing your own voice" in translations of your own writings.

Looking at the broader pedagogical implications of GenAI most of the tested tools could better be used in brainstorming activities and preparation. The very nature of GenAI is to collect and present what already is known with the possibility to select and aggregate from huge amounts of data. However, since there also are several recent research reporting on the darker side of GenAI with issues such as bias and deepfakes (Cooper and Tang; Ferrara, 2024), it seems important, as suggested by Van Dis et al. (2023), to hold on to human verification by teachers and subject matter experts. Looking at this from the student perspective the issues of authorial identity of solutions and accountability are essential (Duah and McGivern, 2024). On the brighter side of these new conditions is that, it could act as a catalyst for the revision of traditional assignments that only involves the lower levels of Bloom's taxonomy (Figure 2).

3.2 Assignment Complexity

As mentioned in the previous section more complex assignments involving the higher levels of Bloom's taxonomy were more difficult to solve correctly. An example that is investigated and discussed in one of the group reports is a two-part assignment on document handling plans. In part one of the assignment the students should describe and discuss what document handling plans are, what they could help out with, what they could be used for, how useful they are, and what pre-knowledge that is required for using document handling plans. Part two of the assignment is about a specific area, where specific document handling plans are given, and where students should discuss and reason around the added value of document handling plans. Two web resources that contain content to answer these assignments are given, and that they both should be referenced to in the solutions.

Students asked a GenAI tool to generate the 2000-word report that was required in the assignment, based on the document handling plans for two Swedish authorities (The Ministry of Justice and The Taxing Authority). The generated output consisted of 485 words where the prompted questions only were partly answered, and without any references at all. In another prompt, the chatbot was asked to elaborate, and to add the required references. Some more information was added, but not much and this type of assignment structure must be seen as 'AI resilient'. When Bard (now Gemini), was prompted another time to elaborate and to add references, the software completely crashed. In the detection software Smodin, the 458 words were judged as AI generated with a probability of 71.8 %. Moreover, the solution of part 2 was also incomplete and the 537 generated words were evaluated as " Human and AI generated" with a probability of being 60.8% AI generated. All group reports criticize the output from the tested detection software which bridges to the theme of the next section.

3.3 Detection Software

Due to the limited course budget students had to use detection software without any licence fees. Some of them performed relatively poor and one group report states that "According to our tests Copyleaks had a higher degree of correctness than GPTZero". Where Copyleaks at that time had an estimated accuracy rate of 66% (Figure 1). The same group report also indicates that human generated texts with spelling errors and grammatical flaws more easily pas the AI detection. Not a surprise, and what was more surprising was a teacher answer to an interview question about the use of AI detector software: "I don't really understand the question if it considers student texts in general, or specifically AI generated texts. Student texts are always run in plagiarism checking software. If AI generated texts get detected more frequently, I don't really know". What this teacher, and many other teachers hope for is that the AI detection should be a built-in part of the plagiarism

control, which now is the case in systems such as Turnitin. However, even if Turnitin has a higher accuracy rate than most of the software listed Figure 1, the accuracy rate was between 75-85% when tested. Where Turnitin stands out is regarding false positives where the test result was 0.00% (Weber-Wulff, 2023)

Another group made tests with the detection software tools Smodin and Copyleaks, where they found both false positives and false negatives. "The Smodin tool showed variations in the evaluations between different executions of the same text, and evaluated large part of an AI generated text as human generated, and the same for Copyleaks". The same group wrote in their report, "this aligns with the conclusions in Farrelly and Baker (2023), that is today not possible to trust the detection tools. There are too low accuracy rates for using the evaluations as proof in the judgement of student submissions". The same recommendation is also given by Tate et al. (2023), where this study points out that the current detection tools are not trustworthy when it comes to decisions about if a text is generated by a human or by an AI tool. The third group made tests with ZeroGPT and found accuracy rates between 46.6 - 67.7% for AI generated texts, to compare with the 99% that can be found on the providers web page. In the middle of the course span the university had the first case of suspected AI generated student submissions, where one student was convicted in the university discipline board. In media it was stated that the conviction was based on an evaluation with an AI detection software. Two of the group reports posit their doubts about using detection software in this kind of cases.

3.4 Ethical Issues

In the group reports the most frequently mentioned issue is about detection software, and the lack of trust in detection software. Several students use the Swedish word "rättssäkerhet", a term that DeepL translates to legal certainty or legal security, but a term that in Swedish also involves transparency and citizens trust of a legal process. In the individual essays several students bring up the question of that "GenAI could be misused and that students then don't acquire the skills and knowledge that are stated in the learning objectives". However, in the same essay where the quote above was taken from it was also questioned "Is it that wrong to use GenAI in studies?" and "Isn't AI also a knowledge in itself?". Another essay brings up the ethical issues of Deepfakes, and the importance of keeping higher education free from Deepfakes. This essay also highlights how GenAI can be a double-edged sword in programming education, generating code that the students cannot understand. Moreover, this essay also points out the problems with: "GenAI substituting teachers", and "the ownership of AI generated content".

One essay that suggests that assignments that involves more of practical real-world problems would be harder to GenAI solve, also compares GenAI to other technologies. Like Einstein said about nuclear energy, it can be used both for good and for bad. Another essay that brings up the issue of "chatbot hallucinations" also discusses the issue of biased training data, and that chatbots only generate content based on what their neural networks have been fed with. Other essays discuss the issue of integrity, and how much, and which personal information that is stored on different Internet servers. One student wrote the essay partly around a conversation with a chatbot, where the chatbot was asked to bring up ethical issues. In the reply there were three issues that also could be found in several other essays: bias, lack of transparency and the earlier mentioned integrity. This also align with the findings in Zhai (2022), a study where ChatGPT was prompted to write a paper about challenges with GenAI in educational contexts.

3.5 Recommendations for Redesign

One of the groups created a list of recommendations for subject matter experts and instructional designers that is worth considering:

- Encourage students to use GenAI tools to explain complex concept, but that their submissions must be complemented with references to other verifying sources
- Check all the references in essays and other textual submissions
- The textual assignments should be complemented with oral examination
- Test all assignments on GenAI tools before giving them to students to evaluate how easily they can be solved with AI support
- Create assignments where the analysis should be filtered by the course literature and involve own opinions
- Continue to use traditional methods such as to discuss the use of AI with students, instead of only using detection software
- Require students to declare if GenAI tools have been used in a solution, and which tool and in which parts of the solution

- Involve assignments where students should find innovative solutions to real-world problems in areas where students are supposed to work in the future
- Encourage students to interdisciplinary research on AI, technology and education, and that researchers in higher education should have a dialogue with companies and organisations that develop GenAI tools

The last point in the list has its origin in the group' s report from the first Symposium on AI Opportunities and Challenges that was a mandatory part of the course (Jaldemark, Mozelius & Humble, 2023). Another group has similar suggestions and sees oral examinations as a kind of active learning events that could reinforce the learning outcomes. Moreover, they do not want to ban GenAI tools, but they want "clear guidelines and a university policy that states when and where AI tools are permitted". The third group also has similar recommendations about new examination forms, new ethical guidelines, and that the university should evaluate the quality of detection tools.

4. Concluding Discussion and Future Research

Data were collected from students examining their own Master's program which could introduce bias, as students may have preconceived opinions about the program and its assessments. This could affect their evaluations and the study's outcomes as in all educational action research studies where data are collected by researchers with dual roles. Moreover, the relatively small sample size limits the generalizability of the findings to a broader population of students or other educational programs. This study should be seen as a first explorative step in the educational action research process that most departments need to carry out to redesign courses for a higher GenAI resilience.

Some course related ethical aspects were brought up from the student perspective in the result section, but there are also broader ethical implications and other stakeholder involved in the use of GenAI in higher education. As brought up in the study on GenAI research ethics by Van Dis et al. (2023), it could be important to widen the debate to involve more stakeholders, hold on to human verification, and to to develop rules for accountability. Other ethical aspects to tackle include bias in the AI models, stewardship and transparency. In a European context it would also be highly important to provide a data privacy that aligns with the GDPR law (Leboukh, Aduku, and Ali, 2023).

The conclusion resembles the one in Mozelius (2024), a majority of the current assignments in the investigated Master's programme could be partly solved with different GenAI tools. At the same time, the solutions to the more complex assignments showed to be incomplete as well as incorrect. A new finding is that students and teachers have a consensus on that GenAI should not be banned, but that there is a need for a general policy and specific guidelines for all courses in the programme. Moreover, the tested detection software showed poor results leading to the ethical concern of using this kind of software in higher education disciplinary boards. The recommendation is to revise the current assignment design in courses to more varied process that involves more of real-world problems and creativity. The revised assignments should also provide guidelines for which GenAI tools are allowed and when, and on the other hand that students should declare if GenAI tools have been used in a solution, and which tools and in which parts of the solution.

As future research it would be interesting to the presented recommendations for redesign in the development of a new course, and to try to make the course GenAI resilient from start. This should be a course on AI in education where GenAI tools will be used to a relatively high degree. At the time of writing the author has started the development of a new course on AIED where some of the questions that were unanswered in this study will be asked again. Finally, the broader pedagogical implications of involving GenAI in higher education need further exploration.

References

Armstrong, P. (2010). Bloom's Taxonomy. Vanderbilt University Center for Teaching. Retrieved (03/07/2024) from https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/

Arnold, L. and Norton, L., 2021. Problematising pedagogical action research in formal teaching courses and academic development: A collaborative autoethnography. *Educational Action Research*, *29*(2), pp.328-345.

Baum, F., MacDougall, C. and Smith, D., 2006. Participatory action research. *Journal of epidemiology and community health*, *60*(10), p.854.

Chan, C.K.Y. and Hu, W., 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), p.43.

Chiu, T.K., 2024. Future research recommendations for transforming higher education with generative AI. Computers and Education: Artificial Intelligence, 6, p.100197.

Cooper, G. and Tang, K.S., 2024. Pixels and Pedagogy: Examining Science Education Imagery by Generative Artificial Intelligence. *Journal of Science Education and Technology*, pp.1-13.

Duah, J.E. and McGivern, P., 2024. How generative artificial intelligence has blurred notions of authorial identity and academic norms in higher education, necessitating clear university usage policies. *The International Journal of Information and Learning Technology*, *41*(2), pp.180-193

Duc, A.N., 2023. Exploring the Potential Impact of Generative AI in Education–A reflective analysis from nine IT courses. In *Norsk IKT-konferanse for forskning og utdanning* (No. 2).

Farrelly, T. and Baker, N., 2023. Generative artificial intelligence: Implications and considerations for higher education practice. *Education Sciences*, *13*(11), p.1109

Ferrara, E., 2024. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, pp.1-21.

Forbes 2023. "The 10 Greatest AI Content Detector Tools" Retrieved (15/12/2023) from https://www.forbes.com/sites/technology/article/best-ai-content-detector-tools

Jaldemark, J., Mozelius, P. and Humble, N., 2023. Symposium on AI Opportunities and Challenges: Education will never be the same again. ACI Academic Conferences International.

Leboukh, F., Aduku, E.B. and Ali, O., 2023. Balancing ChatGPT and data protection in Germany: challenges and opportunities for policy makers. *Journal of Politics and Ethics in New Technologies and AI, 2*(1), pp.e35166-e35166.

Luo, J., 2024. A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education*, pp.1-14.

Mishra, P., Warr, M. and Islam, R., 2023. TPACK in the age of ChatGPT and Generative AI. *Journal of Digital Learning in Teacher Education*, *39*(4), pp.235-251.

Mozelius, P., 2024. Generative AI and Assessment in Higher Education: The Student Perspective. In *SAIOC 2* (Vol. 2, p. 23). ACI Academic Conferences International.

Mozelius, P., Cleveland-Innes, M., Håkansson Lindqvist, M., Jaldemark, J., and Sundgren, M., 2024. Educating the educators on artificial intelligence in higher education. In *Proceedings of the 4th International Conference on AI Research, ICAIR 2024.* Lisbon, Portugal. Academic Conferences International.

Norton, L., 2009. *Action research in teaching and learning: A practical guide to conducting pedagogical research in universities*. Routledge.

Pons, A., 2023. Generative AI in the classroom: From hype to reality.

Stewart, B., 2023. GenAI on GenAI: Two Prompts for a Position Paper on What Educators Need to Know. *Irish Journal of Technology Enhanced Learning*, *7*(2), pp.32-41

Tate, T., Doroudi, S., Ritchie, D., Xu, Y. and Warschauer, M., 2023. Educational research and AI-generated writing: Confronting the coming tsunami. *EdArXiv. January*, *10*.

Van Dis, E.A., Bollen, J., Zuidema, W., Van Rooij, R. and Bockting, C.L., 2023. ChatGPT: five priorities for research. Nature, 614(7947), pp.224-226.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. and Waddington, L., 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), p.26.

Whitham, R., Stockton, G., Richards, D., Lindley, J., Jacobs, N. and Coulton, P., 2023. Implications of Generative AI on Learning and Assessment in Higher Education and Design Research Practice.

Zhai, X., 2022. ChatGPT user experience: Implications for education. *Available at SSRN 4312418*.

Educating the Educators on Generative Artificial Intelligence in Higher Education

Peter Mozelius[1], Marcia Håkansson Lindqvist[1], Martha Cleveland-Innes[1,2], Jimmy Jaldemark[1] and Marcus Sundgren[1]
[1]Mid Sweden University, Sundsvall, Sweden
[2]Athabasca University, Alberta, Canada

Peter.Mozelius@miun.se
Marcia.HakanssonLindqvist@miun.se
martha.clevelandinnes@miun.se
Jimmy.Jaldemark@miun.se
Marcus.Sundgren@miun.se

Abstract: In the current spring of Artificial Intelligence, the rapid development of Generative AI (GenAI) has initiated vivid discussions in higher education. Opportunities as well as challenges have been identified and to cope with this new situation there is a need for a large-scale teacher professional development. With basic skills about GenAI teachers could use the new technology as an extension of the existing technology enhanced teaching and learning. The aim of this paper is to present and discuss the project FAITH (Frontline Application of AI and Technology-enhanced Learning for Transforming Higher Education). FAITH is a higher education pedagogical development initiative for institutional development for teachers with good fundamental skills in traditional pedagogy. A project with the overall objective of increasing the staff understanding of AI and to develop new competencies in the field of GenAI and technology enhanced learning. The research question that guided this study was: "What are the perceived opportunities, challenges and expectations of involving GenAI in higher education?" The overall research strategy for the FAITH project is design-based research, which involves iterative and cumulative development processes. In the early iteration that this study was a part of has been carried out inspired by Collective Autoethnography where members of the steering group behind the FAITH project, and members of the project team have constituted the main focus group. Data were collected by structured interviews where two GenAI tools also have been interviewed. Findings show that the expectations are high, but that the FAITH ambition of institutional development is depending on teachers' motivation for taking an active part in the project. Another challenge could be that many teachers see GenAI as something that threatens the current course design, and that a general ban of GenAI is the appropriate solution. One of, several identified opportunities, is that a general revision of syllabi and assessment in an adaptation for GenAI enhanced learning would improve the current course design.

Keywords: Teacher professional development, Artificial intelligence in education, AIED, GenAI, Higher education, Transformation of higher education

1. Introduction

Current narratives about AI in education cover the benefits, the challenges, and the need for ethical practices. The suggestions that beneficial AI can handle interpretation of very large data sets and offer personalized learning is noted. However, the high-quality datasets that are the substance of GenAI are labelled by humans and carry ontological assumptions that may or not be shared. (Vuong et al., 2018). To deal with current and emerging AI in education, we must "1) demystify new AI systems; 2) focus more on the mundane technical problems of reliability/accuracy, intrinsic biases, vulnerability to hacking, etc., rather than the threat of the "superintelligent AI," and 3) promote balanced ethical discussion about both the potential benefits and harms of these systems among all stakeholders" (Tung, 2020, p. 4). These issues will shape the content and dialogue for FAITH participants.

Current narratives about AI in education cover the benefits, the challenges, and the need for ethical practices. The suggestions that beneficial AI can handle interpretation of very large data sets and offer personalized learning is noted. However, the high-quality datasets that are the substance of GenAI are labelled by humans and carry ontological assumptions that may or not be shared. (Vuong et al., 2018). To deal with current and emerging AI in education, we must "1) demystify new AI systems; 2) focus more on the mundane technical problems of reliability/accuracy, intrinsic biases, vulnerability to hacking, etc., rather than the threat of the "superintelligent AI," and 3) promote balanced ethical discussion about both the potential benefits and harms of these systems among all stakeholders" (Tung, 2020, p. 4). These issues will shape the content and dialogue for FAITH participants.

FAITH is a project where the overall objective is to increase the staff understanding of AI and to develop new competencies in the field of GenAI and technology enhanced learning. The main research question to answer in

this study was: "What are the perceived opportunities, challenges and expectations of involving GenAI in higher education?". A fundamental idea in the FAITH project is to explore and analyse the capacity of GenAI tools instead of rejecting them. Some researchers see GenAI chatbots as co-researchers today that can do a wide range of research activities (Zhang et al., 2023; Barany et al., 2024), but in this study chatbots have only been a minor part of the data collection.

2. The FAITH Project

Higher education and technological development have a symbiotic relationship. On one hand, much technological development is a result of research efforts. In turn, these efforts lead to technological applications in society. One such field is digital technologies. This is a field, which, nowadays, is very focused on AI technologies. On the other hand, these technologies are applied in trials that aim to develop higher education operations. For example, in how AI impacts learning and teaching in higher education. This in turn is the focus of applied research, for example in the field of educational science. This paper illustrates this process by discussing a development project called FAITH (Frontline Application of AI and Technology-enhanced Learning for Transforming Higher Education).

FAITH is a professional development project that runs between 2024-2026 and aims to develop institutional teaching development and education programs further. In the project, teachers at the studied university will work and develop its educational operations to adapt it to work and study in a highly communicative, cloud-based, mobile and smart AI-based society. The integration of AI and technology-enhanced learning is a key to achieve this goal. The FAITH project will take advantage of ongoing higher education pedagogical development initiatives within the university; harness the power and benefits of already established models and teams, develop professional knowledge through new higher education pedagogy; and increase staff understanding and competence in AI and technology-enhanced learning (Jaldemark et al., 2024).

In professional development efforts, design-based research is a key methodology which involves iterative and cumulative development processes (e.g., Jaldemark, Lindqvist, & Mozelius, 2019; Jaldemark & Öhman, 2020). FAITH consists of three phases: preparation, implementation and evaluation. The first phase is preparation, which includes detailed planning, forming a steering group, and recruiting key competencies for the group team. The project team consists of AI and technology-enhanced learning experts and program managers. These groups are responsible for the staffing and scheduling of FAITH, including developing a plan for the competence development action. Program managers have a prominent role in the FAITH project team and are responsible for the pedagogical development of their respective programs. In the second phase, the implementation phase, the project team supplies the teachers with AI and technology-enhanced learning skills training, supporting pedagogical development work and dissemination of ideas through seminar activities. The focus is on the impact of AI and technology-enhanced learning on higher education studies, including both AI as content and form. The third phase, the evaluation phase, compiles experiences and conclusions are drawn. This phase also proposes further work and ideas that should be implemented in the coming iterations of the included courses and programmes.

3. Related Research

During the last years there has been a fast development of new tools that build around large language models and GenAI techniques. This started with the release of ChatGPT in late 2022 with a development of various GenAI tools that has created both promising opportunities and serious challenges in higher education (Chan and Hu, 2023; Yeralan and Lee, 2023). Many research studies have reported on the need of updated assessment methods to secure that students submissions are true and original work (Gamage et al., 2023; Luo, 2024). The ongoing AI hype might be a hype followed by a new so called AI winter, but the strong impact on teaching and learning in higher education will remain as a challenge to address (Pons, 2023; Whitham et al., 2023). Several universities have rapidly created draft policies that sometimes contradict each other, while other universities at the same time have kept a low profile and remained silent. As highlighted in the study by Luo (2024), to strict and banning policies could result in anxious students that hesitate to use GenAI tools even for legitimate learning purposes.

Higher education and technological development have for many years had a symbiotic relationship that has created bot opportunities and challenges. On one hand, new technologies could be seen as a result of research efforts that has led to technological applications in higher education like the currently strong focus on GenAI technologies. On the other hand, these technologies have also been applied in projects that has aimed to further develop higher education pedagogy and to investigate how new technology might support learning and teaching

in higher education. (Mozelius et al., 2024) As pointed out by many recent studies, there is a strong need today for higher education initiatives on how to involve GenAI in course design and students' learning activities, and for teacher professional development on GenAI (Brandão, and Zagalo, 2024; Meli, Taouki and Pantazatos, 2024; Mozelius, 2024; Mozelius et al., 2024).

4. Method

In the described FAITH project, the overall research strategy is design-based research, which involves iterative and cumulative development processes as described by Hoadley and Campos (2022). This early iteration that this study is a part of was carried out inspired by Collective Autoethnography (CoAE). Authors have followed the six-phase approach outlined by Karalis Noel, Minematsu and Bosca (2023), that involves 1) Preparation, 2) Data collection, 3) Transcription, 4) Interpretation, 5) Thematic consensus, and finally, 6) The narrative production. CoAE is an approach that aligns with the spirit of participatory research, since it involves all participants collectively throughout the entire research process, from the definition of the research question to the writing of the final narrative (Karalis Noel, Minematsu and Bosca, 2023). As in participatory action research all participants in the research team should be part of data collection and data analysis, with a focus on collaboration mutual learning, and co-creation of knowledge (Greenwood and Landevin, 2006; Kindon, Pain and Kesby 2007; Karalis Noel, Minematsu and Bosca, 2023).

The members of the FAITH steering group and members of the project team have constituted the main focus group, as well as the author team. In the first preparation phase, authors discussed the study design, the research question and how to collaborate. The two following phases were carried out collaboratively with data collection by structured interviews where members of the focus group and two GenAI tools have been interviewed. To use chatbots in research is not a new concept, and was tested by Shawar and Atwell (2007), many years before the appearance of GenAI bots. Today, researchers such as Barany et al. (2024) consider chatbots to be both tools and co-researcher, and as stated by Zhang et al. (2023) chatbots could act as co-researchers and assist, challenge, or supplement the human researchers. In this study the chatbots have only had a minor role in the data collection, and all the important analysis and synthesis were done by the authors. The two used GenAI tools were ChatGPT and Gemini, that were asked the same questions as the human members of the focus group about the FAITH project. In the fourth phase of review and naming themes, all interview answers were thematically analysed by 2 focus group members according to the process defined by Braun and Clarke (2019). This was followed up in the fifth phase by discussions among all focus group members to reach a thematic consensus before the final collaborative narration production.

5. Findings and Discussions

In a deductive thematic analysis, the answers have been grouped into the predefined categories of 'Expectations', 'Challenges' and 'Opportunities'. These categories are presented and discussed one by one here below with a division between human answers and GenAI chatbot answers. Humans are referred to as Informant 1-5, and the GenAI tools as Chatbot 1 (ChatGPT 4) and Chatbot 2 (Gemeni Basic).

5.1 Expectations

Both humans and chatbots show a general positive attitude toward the FAITH project where the human informants have the more detailed information. Chatbots only had the project description that can be found before the method chapter as input.

Human informants

Informant 1 wrote that "I have very high expectations for the FAITH-project", and that the Department of Education needs to be in the forefront in the AI field. As for many other departments it is both about supporting students, and to support teachers design of AI-based teaching and learning activities. If thoughtfully implemented, this could be "a collective knowledge boost for the whole institution" (Informant 3), and that it also could be seen as "a valuable extension to the previous work on higher education reform and lifelong learning completed by this research team" (Informant 2). The same informant, who might be the one with the highest expectations, also foresees a future with increased "professional knowledge through new higher education pedagogies" at the institutional level. In combination with the micro-level individual learning that will provide an "increased staff understanding and competence in AI and technology-enhanced learning".

Moreover, there are expectations of a research part of the project where Informant 4 describes the project as an interesting "case study of AI applications in higher education", where it will be possible to "observe students

and teachers' applications of AI and how it impacts their learning and higher educational setting". All informants mention the internal knowledge sharing, while Informant 3 also brings up that "The results we achieve locally should be discussed internationally in an exchange of 'lessons learned' and 'best practices'". In the answer from Informant 5 there are expectations about beneficial outcomes in at least the three areas of: "1) discovering relevant applications or practices that can be taught to our students, 2) discovering relevant applications or practices that can assist in their own teaching, and 3) discovering relevant applications of GenAI that become helpful in their personal administration, preparation, and planning".

GenAI Chatbots

Chatbot 1 is as optimistic as the humans with high expectations in the areas of 1) Enhanced Educational Operations, 2) Professional Development, 3) Leveraging Existing Initiatives, 4) Iterative Development, and 5) Phased Implementation. Iterative development is built on the use of design-based research methodology in the project description, where iterative development is a corner stone in design-based research projects that Chatbot 1 also mentions that "the project expects to engage in iterative and cumulative development processes to refine and enhance educational practices continuously". Which is exactly according to the project plan, and the same for the elaboration of Phased implementation: "The project is expected to follow a structured approach with clear phases (preparation, implementation, and evaluation) to ensure systematic development and assessment of the project's impact on higher education".

While Chatbot 1 reformulates and adds information about expectations, Chatbot 2 gives more of a summary referring to the references in the given project description:

5.2 Expectations

- Develop institutional teaching development and education programs further (Jaldemark et al., 2024)
- Adapt the university's educational operations to a highly communicative, cloud-based, mobile and smart AI-based society (Jaldemark et al., 2024)
- Increase staff understanding and competence in AI and technology-enhanced learning (Jaldemark et al., 2024)

Complemented with the positive, but very general "The FAITH project has the potential to significantly improve the quality of education at the studied university by developing new teaching methods and technologies".

5.3 Challenges

With the overall high expectations in the previous section, it is important to look at if there are any perceived challenges that could hamper the project implementation. What was an interesting finding is how the answers from the chatbots converge with the human concerns regarding the challenges of "Ethical Considerations" (Chatbot 1) and "Encouraging faculty participation and adoption of new methods" (Chatbot 2).

Human Informants

In human words the challenge of involving all stakeholders was expressed as to "involve all project members so that they participate fully" (Informant 3) and to "create communication and development plans for the FAITH team members and key stakeholders to ensure project momentum and support" (Informant 3). Regarding what the chatbots mention as "resource allocation" Informant 1 is worried about "that this project will take place at the same time as ordinary work with teaching and research. It might be difficult for teacher and researchers to prioritize the project. This could be due to time and possibility to meet in work groups".

Informant 5 points out the issue with that it is a moving target where "GenAI technology evolves rapidly, which could make it difficult to keep up with the latest when planning and preparing content", and an immature field where "There are still lots of juridical blind spots and grey areas that could prove difficult to navigate. For instance, the EU GDPR regulation could hinder the implementation of compulsory GenAI-based course tasks, as students can’t be forced to volunteer personal data associated with their user account". Informant 5 highlights the complexity of the topic compared to earlier technology enhanced learning projects since "AI have the potential to deeply impact society. Compared to video conferencing, AI is more similar to the implementation of Internet in educational settings. AI cannot as part of our teaching only be a privilege to a certain category of students, they all need to understand its impact on humans and society".

GenAI Chatbots

Both the chatbots list some challenges that are worth considering. Chatbot 1 gave the more elaborated warning with: 'Change Management' explained as the challenge of "Managing resistance to change among faculty and staff who may be sceptical or hesitant to adopt new technologies and teaching methodologies", and 'Technical Expertise' consisting of "Recruiting and retaining experts in AI and technology-enhanced learning who can effectively train and support the teaching staff". Chatbot 2 had a more parsimonious answer with the brief listing of:

- Integrating AI and technology-enhanced learning effectively
- Encouraging faculty participation and adoption of new methods
- Evaluating the impact of AI on learning outcomes

5.4 Opportunities

There were clearly more opportunities in the informants' answers than challenges, which the authors find to be a good sign in the inception phase of a long-term project.

Human informants

Informant 1 mentions opportunities with the FAITH project such as "It places the department in the forefront at the university in taking on the issue of AI", and that "It will provide teacher the opportunity to gain new knowledge in how to integrate AI in their teaching". This resembles the answer from Informant 4 who thinks that the "Opportunities with an intervention project like FAITH is to raise the bar regarding higher education teachers expected knowledge level about AI in education". Informant 2 claims that a thorough implementation of the FAITH project would meet the need of "The need for higher education pedagogical development is both a requirement and a side benefit of examining the challenges and benefits of AI in education, and that "The training of behavioural scientists, teachers, and educators with high digital competence allows for making informed choices".

Informant 5 brings up the more detailed opportunities of "using GenAI agents to train students in qualitative interview techniques", and that "GenAI could turn out to be helpful in handling everyday teaching tasks, such as feedback support". Moreover, teacher could get help with time-consuming tasks such as subtitling recorded lectures for accessibility, and that researchers can use GenAI tools for transcribing recorded interviews (Informant 5). On a more general level Informant 3 points out "That everyone involved in the project will feel more comfortable using different GenAI tools in the future", and "That critical thinking about AIED is strengthened, and that a common exchange of knowledge and skills in the style of a 'Community of Practice' also leads to new insights into teaching, didactics and course design".

GenAI chatbots

As for the other categories earlier, the chatbot summarises some of the more general ideas that humans have brought up. Chatbot 2 have very general formulations such as that the FAITH project could "Improve the quality and effectiveness of higher education", while Chatbot 1 is a bit more specific in "Providing teachers with AI and technology-enhanced learning skills can lead to significant professional growth and open up new career opportunities". Chatbot 2 also mentions the opportunity of preparing students for their future work life, and that Chatbot 1 highlights the opportunity of that "The integration of AI can lead to the development of innovative teaching practices that enhance student engagement and learning outcomes". Finally, the opportunity that both chatbots brings up, and the humans have omitted, is the possibility of personalised learning tracks for students.

6. General Discussion

If the current AI spring is a hype or not will be revealed during the coming year. At the time of writing the share prices are falling for tech companies such as Nvidia, an example of a company that is heavily depending on GenAI. Several experts have also highlighted the more general economic bubble that could burst in the field of AI. However, independent of the financial development in the near future the challenge for the educational sector will remain with several important questions to address. Some examples here are:

- To what degree, and when should students be allowed to use GenAI in their studies?
- How could teachers redesign their courses to involve AI tools in a constructive way?
- How should the evaluation of the impact of AI on learning outcomes be carried out?
- In what ways must assignments be redesigned to keep fairness and quality learning?

Questions that also could act as a catalyst for the redesign of higher education that with or without AI has to be done. The emerging trend of a would certainly initiate a shift from assessing the product to assessing the process where oral examination can have a renaissance. Moreover, with the idea from computer chess, that man and machine is stronger than just one of them (Kasparov, 2007), the current learning activities could be reinforced and involve an AI boosted creativity. Even if the current AI hype would fail, and whatever teachers' and researchers' opinions on GenAI, higher education needs to address the questions listed above. Authors recommendation is broad professional development initiatives with the aim of AIED with a human touch. Five other priorities that will be presented in one of the first FAITH workshops are the ones outlined by Van Dis et al. (2023): 1) To hold on to human verification, 2) How to develop rules for accountability, 3) Investing in truly open LLMs, 4) Embrace the benefits of AI, and 5) To widen the debate.

7. Conclusion

The conclusion is that the expectations are very high, and maybe too high, and that there are more opportunities mentioned than challenges. The authors see this as a good sign, since a project without optimism in the inception phase would raise a lot of question marks. On the other hand, the found challenges will be the most important input to the FAITH project, with a common list of pitfalls to avoid. These pitfalls, are however, common pitfalls in many projects in academia. Time and resources are traditional challenges in context. At the same time, it appears that the optimistic view, will strengthen and uphold the FAITH-project over time. This paper contributes with ideas for a nuanced and realistic implementation of lifelong learning initiatives and teacher professional development courses. Important to analyse and discuss the opportunities and challenges in the current AI hype to avoid the pitfalls with a superficial instructional design. Finally, this paper describes the importance of a continuous lifelong learning in a digital era with rapidly emerging new technology.

Finally, in the third spring of AI in general and GenAI in particular it is hard to foresee how the different AI techniques depicted in figure 1 will develop in the near future. Independent of how the AI hype cycle develops in general, AI will have a substantial role in education and as described in Pons (2023), the current AI hype should hopefully lead to realistic improvements in teaching and learning.

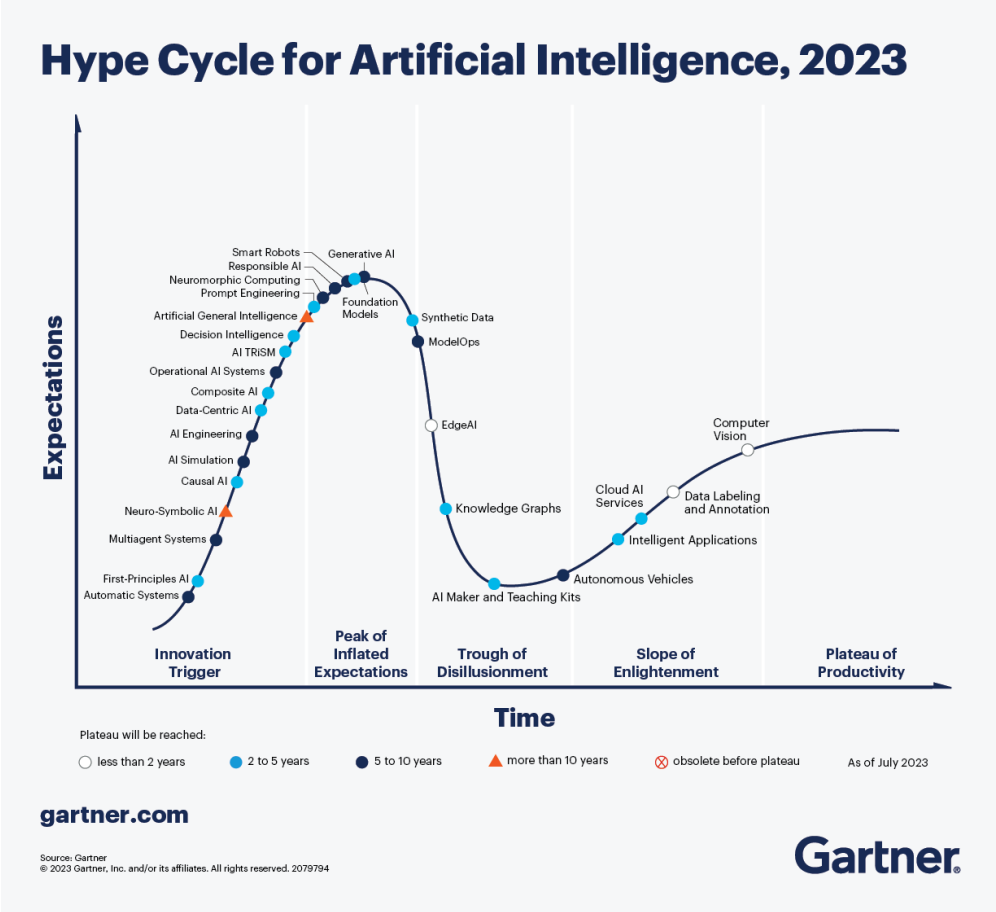


Figure 1: The Gartner Hype Cycle for Artificial Intelligence (Gartner.com, 2024)

References

Barany, A., Nasiar, N., Porter, C., Zambrano, A.F., Andres, A.L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J. and Mehta, S., 2024, July. ChatGPT for education research: Exploring the potential of large language models for qualitative codebook development. In International Conference on Artificial Intelligence in Education (pp. 134-149). Cham: Springer Nature Switzerland.

Brandão, A., Pedro, L. and Zagalo, N., 2024. Teacher professional development for a future with generative artificial intelligence–an integrative literature review. *Digital Education Review*, (45), pp.151-157.

Braun, V., and Clarke, V. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Chan, C. K. Y., and Hu, W. 2023. Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43. https://link.springer.com/article/10.1186/s41239-023-00411-8

Gamage, K.A., Dehideniya, S.C., Xu, Z. and Tang, X., 2023. ChatGPT and higher education assessments: more opportunities than concerns?. *Journal of Applied Learning and Teaching, 6*(2).

Gartner.com, 2024. What's New in Artificial Intelligence from the 2023 Gartner Hype Cycle. https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle (accessed 30/06/2024)

Greenwood, D.J. and Levin, M., 2006. *Introduction to action research: Social research for social change*. SAGE publications

Harry, A., 2023. Role of AI in Education. *Interdisciplinary Journal and Humanity (INJURITY)*, *2*(3), pp.260-268.

Hoadley, C. and Campos, F.C., 2022. Design-based research: What it is and why it matters to studying online learning. *Educational Psychologist*, *57*(3), pp.207-220.

Jaldemark, J., Lindqvist, M. H., & Mozelius, P. (2019). Teachers' beliefs about professional development: Supporting emerging networked practices in higher education. In A. Littlejohn, J. Jaldemark, E. Vrieling-Teunter, & F. Nijland (Eds.), Networked professional learning: Emerging and equitable discourses for professional development (pp. 147–164). Springer.

Jaldemark, J., & Öhman, P. (2020). Developing a hybrid and networked educational approach to lifelong learning for organisations and employees. In M. de Laat, T. Ryberg, N. Bonderup Dohn, S. Børsen Hansen, & J. J. Hansen (Eds.), Proceedings for the 12th International Conference on Networked Learning (Vol. 12, pp. 47–50). Aalborg university.

Jaldemark, J., Cleveland-Innes, M., Mozelius, P., Håkansson Lindqvist, M., & Sundgren, M. (2024). FAITH: Educating the educators on artificial intelligence in higher education. In J. Jaldemark, P. Mozelius, N. Humble, & P. Griffiths (Eds.), *Book of abstracts for the 2nd Symposium on AI Opportunities and Challenges: An avalanche of AI to radically change society* (vol 2., pp. 15–16). Mid Sweden University, University & EM Normandie Business School.

Karalis Noel, T., Minematsu, A. and Bosca, N., 2023. Collective autoethnography as a transformative narrative methodology. *International Journal of Qualitative Methods*, *22*, p.16094069231203944.

Kasparov, G., 2007. *How life imitates chess*. Random House.

Kindon, S., Pain, R. and Kesby, M., 2007. Participatory action research approaches and methods. *Connecting people, participation and place. Abingdon: Routledge*, *260*.

Meli, K., Taouki, J. and Pantazatos, D., 2024. Empowering educators with generative ai: The genai education frontier initiative. In *EDULEARN24 Proceedings* (pp. 4289-4299). IATED.

Mozelius, P., 2024. Generative AI and Assessment in Higher Education: The Student Perspective. *In SAIOC 2* (Vol. 2, p. 23). ACI Academic Conferences International.

Pons, A., 2023. Generative AI in the classroom: From hype to reality.

Raj, R. and Kos, A., 2023. Artificial Intelligence: Evolution, developments, applications, and future scope. *Przeglad Elektrotechniczny*, *99*(2).

Shawar, B.A. and Atwell, E., 2007. Chatbots: are they really useful?. *Journal for Language Technology and Computational Linguistics, 22*(1), pp.29-49.

Tung, H.M., 2020. Five takeaways from Mitchell's Artificial Intelligence: A guide for thinking humans. https://www.researchgate.net/profile/Tung-Ho-3/publication/346996546_Five_takeaways_from_Mitchell's_Artificial_Intelligence_A_guide_for_to_thinking_humans/links/62c3ca2d9a17145f5f4681b7/Five-takeaways-from-Mitchells-Artificial-Intelligence-A-guide-for-to-thinking-humans.pdf

Vuong, Q.-H., La, V.-P., Vuong, T.-T., Ho, M.-T., Nguyen, H.-K. T., Nguyen, V.-H., .Ho, M.-T. (2018). An open database of productivity in Vietnam's social sciences and humanities for public use. *Scientific Data, 5*(1), 180188.

Van Dis, E.A., Bollen, J., Zuidema, W., Van Rooij, R. and Bockting, C.L., 2023. ChatGPT: five priorities for research. *Nature*, *614*(7947), pp.224-226.

Whitham, R., Stockton, G., Richards, D., Lindley, J., Jacobs, N. and Coulton, P., 2023. Implications of Generative AI on Learning and Assessment in Higher Education and Design Research Practice

Yeralan, S., and Lee, L. A. 2023. Generative AI: Challenges to higher education. *Sustainable Engineering and Innovation*, *5*(2), 107-116.

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J. and Carroll, J.M., 2023. Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv preprint arXiv:2309.10771*.

Transcending Fixed Meanings: Exploring the Impact of Linguistic Relativism on Adaptive Language Models in Generative AI

Subhangi Namburi
University of Nottingham, UK

subhangidasnamburi@gmail.com

Abstract: This research paper aims to start a discourse exploring the impact of linguistic relativism on adaptive language models within the field of generative AI, challenging the traditional fixed-meaning approach to natural language processing (NLP). It argues for a shift towards more personalised AI systems that can adapt to individual users' language nuances, rather than relying solely on large datasets with predetermined meanings. The current NLP models, based on conventional semantics, assume that language has a stable, objective reality where words have universally accepted meanings. This approach limits AI's ability to understand and generate language that reflects personal or contextual variations. The paper debates that generative AI should move towards a model that embraces the fluidity and subjectivity of language, where meanings are not fixed but can change depending on the speaker's intent or the situational context. This would involve incorporating user-specific data and situational awareness into AI systems, enabling them to interpret not just the literal meanings of words but also the speaker's intentions and the circumstantial cues that may alter these meanings. Such an approach would lead to the development of AI systems that are more adaptive and sensitive to the nuances of personal expression and contextual interpretation. However, the paper also acknowledges the potential ethical challenges associated with this approach. If AI systems are designed to allow for fluid and personalized meanings, they could be manipulated to shape public discourse in ways that reflect the biases or intentions of their developers. This raises concerns about the potential misuse of AI in influencing perceptions and realities, particularly when the fluidity of language is taken to an extreme where communication becomes chaotic and ineffective. Ultimately, suggesting personalised language models offer significant potential for enhancing AI's ability to understand and generate human-like language, there is a need for a balance between individual linguistic creativity and the communal aspects of language that ensure effective communication. The paper concludes that integrating linguistic relativism into AI models could advance the theoretical understanding of language in AI, but it must be approached with caution to avoid undermining the stability and clarity essential for meaningful human interaction.

Keywords: Linguistic relativism, Generative AI, Adaptive language models, Ethical AI, Language learning applications

1. The Postmodern Reconstructionist Approach to Language Learning

This paper explores how does the integration of generative AI in language learning applications reflect and reshape cultural narratives in language acquisition. By starting a discourse, I seek to delineate both the potential and limitations of AI-driven tools in creating meaningful, contextually grounded language learning experiences.

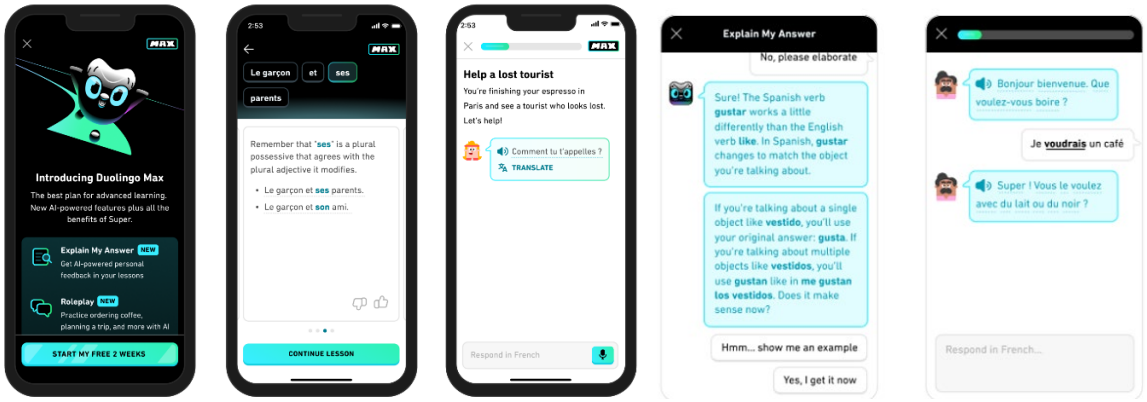
The postmodern reconstructionist approach to language learning debates the goal of language learning being that of transitioning from a Eurocentric 'dominator' model of culture; towards an aesthetic of interconnectedness, social responsibility and cultural attunement. This reconstructionist view of language learning provides optimism for language education by recognising language as existing within the culture, rather than alongside culture. Language then becomes a catalyst of change. The evolutionary perspective on language learning debates how closely intertwined language and culture are as adaptive mechanisms for human survival and socialisation. Language is a catalyst in understanding the cultural, social and political issues individuals face. It aids in coordinating their actions and social bonds. On the other hand culture as a shared system of beliefs, values, practices and traditions provides a sense of identity and belongingness within a group. Cultural practices and beliefs are communicated through language, and language reinforces cultural norms and values transmitting knowledge across generations, allowing cultures to evolve and adapt to changing circumstances over time and survive in a constantly changing environment. Language learning (LL) helps individuals interact with an international community, and interaction as an approach leads to more effective language development. The personal and interactional functions of language help shape the projected identity of an individual, their opinions and preferences and development of social bonds.

The 21st century welcomed a stream of language learning applications (LLA) aiding the reconstructionist approach to LL globally. The acceptance and use of the LLA surged because of the ability to teach and learn languages inter-continentally. This intertwined relationship between language and culture serves as the foundation for developing innovative approaches to language learning, using generative AI being one (Namburi & Hopkins, 2023).

2. Limitations of Generative AI in Understanding Context

In the field of language learning, there has been an ontological shock because of the recent advent of LLMs and generative AI. Generative AI, in the form of chatbots has penetrated into language learning, including those used in Duolingo and Memrise; they are fundamentally limited by their lack of embodiment and contextual awareness. They rely solely on textual or spoken input, missing out on the essential non-verbal cues that play a significant role in human conversations.

Image 1,2 and 3. Use of Generative AI in Duolingo



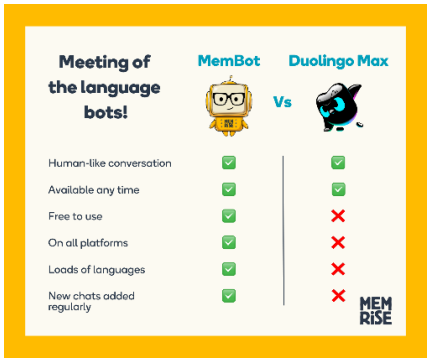


English Translation: Bot: Yes. Do you have pain in your eyes?

Human: No. I cry only on the inside.

Bot: Okay! So, what's up?

Image 4



Memrise's claim of distinction over Duolingo in integrating AI into its software appears to be more a matter of marketing than substantive advancement. The benefits highlighted in their approach primarily reveal a superficial incorporation of AI, rather than a genuinely robust or transformative integration.

As a result, the chatbots' interactions are shallow and lack the depth needed for immersive language practice. They do not possess personal temporality or memory, which is crucial for understanding nuanced, long-term language learning. AI systems cannot retain long-term conversational context, limiting them to brief, turn-based interactions.

Use of language is not merely a process of symbol manipulation. Generative AI systems exist outside of the world, and only have access to data about the world that we provide to it. For this reason, the output of AI systems, even when they are linguistically competent, are ultimately mostly meaning-less and not the product of a true mind as we conceive it. Language and communication are highly dependent on context, and not just on general contexts like register, audience, etc. that could be given in a prompt to an LLM as an initial constraint. There are other contexts like the history you have with someone, previous conversations etc. LLMs are currently limited to a finite number of tokens that keep track of the context of a conversation, and this practical limit constitutes a memory context of about a few minutes in the real-world. We may gradually stretch these contextual limits as AI technology matures, but even under the best of scenarios it will be hard for AI agents to matures beyond a functionally limited conversational utility in the near future (Beals, 2024).

The problem with incorporating Generative AI in language learning applications is that AI relies on inductive reasoning and updates to generate responses. These systems are built on massive datasets and patterns but do not create new knowledge or form genuine explanations. Instead, they are restricted to rearranging pre-learned patterns. When you study and use a language for long enough, and make it a part of your own identity, it starts to change you. You adapt your thought processes to incorporate that language into your actual thinking. It becomes reflexive and natural. From a socio-linguistic standpoint the languages you acquire begin to inform your thinking and being as you navigate the world not just on a conscious level but also on the level of pre/unconscious thought. We develop shorthand modes of operation based on many assumptions of our world, including how we describe and explain it linguistically. Learning new languages is one way we have to develop new modes of being and understanding the world around us. Much of what human thought consists of on a moment-to-moment basis relies on pre/unconscious processes that, when functionally replicated by AI become highly resource-dependent and ungeneralisable to other domains.

This research paper hence aims to open a discourse about the practical implications and challenges of applying a flexible, personalised approach to language in the field of natural language processing (NLP) and generative AI and potentially influence both theoretical and practical advancements in the field. AI systems are designed to interpret and generate language based on fixed meanings derived from large datasets, which may not account for the nuances of personal or situational language use. This paper suggests a shift towards more personalised AI systems that can adapt to individual users' language nuances, challenging the current one-size-fits-all approach in AI.

3. AI Systems' Limitations in Language Acquisition

When AI chatbots generate sentences or content in LLAs, they do so based on statistical likelihood rather than a real understanding of context or user-specific learning needs. This limitation hinders the AI chatbots' ability to offer meaningful, adaptive learning experiences that go beyond surface-level interactions. Although chatbots simulate real-world conversations, the depth of understanding required to foster genuine language proficiency is still missing, making these interactions feel robotic and disconnected from the learner's actual progress or personal experiences. AI can now handle in-distribution data, but it still fails when encountering out-of-distribution situations, for example: when a learner brings up an unfamiliar cultural reference or uses language in a non-standard way, AI chatbots struggle to interpret or respond effectively. This makes the conversational practice less adaptive than it might initially appear, particularly for intermediate to advanced language learners who want to use language creatively.

AI systems cannot replicate this, limiting their utility in higher-level language acquisition. An additional concern also being the reliance on AI for learning can create biases and reinforce incorrect assumptions. This could manifest in reinforcing certain patterns or stereotypes embedded in the training data, leading to a less diverse or culturally sensitive learning experience. Furthermore, the lack of explainability in AI decisions—why a chatbot suggests one answer over another—can diminish trust and hinder the user's confidence in the learning process. These limitations of generative AI underscore a deeper issue regarding the understanding and flexibility of language, particularly as learners advance and introduce more creative, context-specific usage (Estes, 2022).

4. The Grounding Problem

Current AI systems, particularly Natural Language Understanding (NLU) models, grapple with the grounding problem, where symbols and language in AI do not inherently have meaning but rely on external interpretations. This critically raises the issue where symbols or words in a language model do not inherently have meaning. Instead, they rely on external interpretations—typically provided by humans or predefined rules—rather than having any intrinsic understanding of what the symbols represent. This becomes problematic when these models are expected to engage in tasks that require more than just pattern recognition or the mimicking of human language structures; they need to "understand" language in a meaningful way (Beals, 2024).

The symbols or words do not point to anything in the real world on their own. Instead, they are manipulated according to statistical patterns learned from large datasets of text. This is akin to handling words purely based on their form (e.g., the sequence of characters in a word) without understanding their underlying content or connection to the real-world phenomena they describe. This leads to the grounding problem—the model lacks a way to connect its internal symbols to the real-world objects, events, or experiences that those symbols are supposed to represent. Without grounding, AI cannot adapt its interpretation of language based on changes in personal context or nuance. Instead, it is locked into predefined associations, which restricts its flexibility and makes it unable to genuinely "understand" language as humans do. AI systems are then stuck at a surface level, manipulating symbols without grasping their deeper significance, thus making them inadequate for true conversational engagement (Estes, 2022).

AI relies on the use of relational knowledge bases, which, from a philosophical perspective, more closely resembles the associationist approach of conceptual structure (set theory). Whilst statistical approaches that train models on the intended result of a given query/command may be easier to train, and more generalisable to different applications, they face their own limitations. That is, they not only face the general challenges of statistical (deep) learning systems, but they are limited in their ability to interpret novel or ambiguous input—given their dependence on familiar examples from a given training set; as well as their lack of integration with real-world (background) knowledge; and commonsense reasoning (based on notions of causation, rather than mere correlation, between input and output).

From a philosophical perspective, AI's reliance on pre-existing patterns mirrors older conceptual frameworks, further limiting its ability to deal with ambiguity and novel contexts in real-time interactions. AI systems in LLAs should embrace linguistic relativism, wherein meaning is not fixed but instead varies according to context, personal intent, and social dynamics. Traditional NLP systems, with their reliance on fixed, objective meanings, inherently fail to capture this fluidity. The AI systems should be developed to interpret language based not just on static word meanings, but on the nuanced and evolving meanings that depend on personal and situational contexts. To overcome these limitations, AI systems should move toward embracing linguistic relativism, which allows for a more adaptable and nuanced understanding of meaning based on context.

To achieve this, AI systems would need to incorporate a deeper level of grounding—they would need to "understand" language not just in terms of statistical associations but in connection to real-world objects, events, and user-specific cues. This would allow AI to generate more context-aware responses and adapt to the speaker’s unique way of using language. AI could further personalise its responses, offering more relevant and meaningful interpretations based on prior interactions; requiring AI to go beyond statistical associations and incorporate real-world knowledge and cues, making its responses more adaptable and contextually aware (Namburi & Hopkins, 2023).

Human testimony is a significant source of authoritative knowledge. AI systems processing and generating personalised language should incorporate individual testimonies; recognising testimonial authority, especially because AI knowledge sources are unavailable, highlighting the importance of context-driven testimony in shaping reliable language use. This notion of linguistic expression is also the source of authoritative knowledge when other more basic sources of knowledge like sense perception and inference are not available. The emphasis on testimony and trustworthiness in developing NLP models is crucial because it adds a layer to AI's potential for understanding language, showing how context-driven knowledge remains crucial (Language and testimony in classical Indian philosophy, 2020).

AI needs to employ a cognitive-functional grammar (also called a usage-based grammar). Language learners do not simply acquire abstract syntactic rules but rather grasp rules in both syntactic and semantic terms, considering patterns of use and communicative function. The current one-size-fits-all AI model overlooks these

crucial nuances, making it necessary to reconsider how AI could evolve to better capture personal and cultural diversity in language.

5. How Personalisation in AI Language Learning Misses the Mark on Cultural Sensitivity

The GPT model, and other transformer-based architectures have started incorporating context and generate more nuanced interpretations of language by considering the surrounding textual context. They do not operate purely on fixed, objective meanings anymore; instead, they adapt meaning dynamically based on the surrounding words and phrases. While these systems (LLA AI Chatbots) handle context better than older NLP models, they still fall short in terms of personal and cultural contextuality. Their understanding remains surface-level, unable to integrate deeply with personal histories, shared experiences, and non-verbal communication cues that are crucial for inclusive, accurate interactions. One of the critical limitations lies in their inability to maintain a continuous, deep understanding of long-term conversational history. These models typically operate within a finite token window, “remembering" a certain number of words or sentences from the current interaction. Once this limit is exceeded, earlier parts of the conversation are essentially forgotten, preventing the system from drawing upon previous interactions in any meaningful way.

AI is still unable to build on past conversations to form a deeper, more cohesive relationship with the user. This limitation hinders language learning, where the progress and learning trajectory of the user depend heavily on continuous, evolving interactions that accumulate knowledge over time. In human conversations, we adjust our speech and responses based on a shared history of interactions—a quality that is currently beyond the reach of most AI models.While memory-augmented models and retrieval-based methods are emerging, where previous conversations are indexed and recalled when relevant, this technology is still in its infancy. Even these systems struggle to distinguish between what information is contextually important to retain and what can be discarded, leading to either overly repetitive interactions or a lack of meaningful recall.

Language is not just a set of symbols and rules, but a cultural practice that reflects shared meanings, social norms, and values. When humans communicate, they draw not only on linguistic knowledge but also on a deep understanding of cultural references, idiomatic expressions, social expectations, and even historical or political context. A sentence can carry vastly different meanings depending on the speaker’s background, their relationship to the listener, and the specific cultural context in which the conversation takes place.

This gap is especially significant when considering language learning applications, where learners are not just trying to acquire vocabulary or grammar, but also to understand how a particular language is used in real-world, culturally new settings. In language learning, a failure to grasp these cultural nuances can lead to misunderstandings, miscommunication, or a superficial learning experience that doesn’t equip the learner to navigate complex social interactions in a foreign language. AI systems, when lacking this cultural sensitivity, risk reinforcing dominant cultural narratives while marginalising or misrepresenting minority cultures. This creates an incomplete and often biased representation of the language being taught.

Our language is influenced by our motivations, past experiences, current emotional states, and specific needs. The absence of personal temporality means that each interaction is treated in isolation, with little regard for the learner's progress over time or their individual learning style.

Language learning apps are trying to address this through personalisation techniques, such as user-specific profiles that store information about the learner's progress. However, these systems still fall short of the adaptive, intuitive teaching methods used by human educators who can infer meaning from subtle cues, respond to emotional states, and build on personal histories to create a more tailored and empathetic learning experience. Without these abilities, AI interactions remain transactional and impersonal, lacking the emotional depth that fosters long-term engagement and motivation in learners.

The reliance on statistical correlation rather than true understanding presents a profound philosophical challenge. These models are not "thinking" in the way humans do; they are merely calculating the most statistically probable sequence of words given a particular input. This lack of true understanding becomes evident when the AI encounters novel or ambiguous situations, particularly when the language used deviates from standard patterns or when cultural or emotional factors play a key role in shaping the meaning. One of my primary concerns regarding the rapid transformation of AI is that if there is meaning, then there are ideas; if there are ideas, then there is thinking; and if there is thinking, there is at least some degree of consciousness, or at least the beginnings of consciousness forming.

In human language learning, much of what we learn is shaped by abstract reasoning, inference, and the ability to create new meanings from limited information. We don't just learn language through repetition but also by developing conceptual understandings that allow us to generalise across different contexts. AI systems, in contrast, remain limited to the scope of their training data. They still struggle to interpret or generate language in ways that go beyond their statistical foundations, making it difficult for them to navigate ambiguity, handle metaphor, or understand language in its more creative and expressive forms.

On the technological front, AI systems would need to evolve to handle personal data efficiently by employing more compact models that can run on individual devices, reducing the reliance on cloud-based solutions and making real-time personalisation feasible. Advances in edge computing and privacy-preserving algorithms (like differential privacy) are crucial for balancing these demands. Progress in token-based memory management could enable AI to handle longer conversation histories without excessive computational overhead, improving the fluidity and continuity of interactions.

6. Navigating the Ethical Risks

The ethical risks of personalised AI systems, in language learning, are significant, systems could exploit personalised AI to influence or manipulate user opinions by tailoring the information or responses provided based on user-specific vulnerabilities or preferences. These systems could also reinforce linguistic or cultural biases by preferentially presenting dominant cultural narratives, while marginalising minority languages or cultural perspectives. Reinforcing the ethical framework in AI design would require safeguards like bias audits, transparent AI development processes, and multicultural training data that accounts for diverse linguistic and cultural perspectives. AI systems could employ a two-tiered language processing mechanism. The first tier would focus on ensuring communicative clarity by adhering to shared language norms that are widely understood within a community or language group. The second tier could allow for personalisation, where the system adapts to individual linguistic styles, preferences, and creativity.

AI could prioritise clarity by relying on communal linguistic standards as a baseline, while personal deviations would be contextually assessed. The system could flag when a personalised phrase is likely to cause confusion or misunderstanding, and either prompt the user for clarification or offer suggestions to rephrase while maintaining their stylistic intent. AI systems could learn over time how to balance these two aspects by tracking conversational success, ensuring that users are able to creatively express themselves without creating frequent communication problems.

7. AI's Role as a Supportive Tool, not a Creator of new Knowledge

The philosophical aspects of generative AI and it's use deserves greater attention and current discourse often neglects AIs identity of "knowledge creation". This deficit is because of philosophical misconceptions about the growth of knowledge generated using generative AI and the mistaken theoretical understand of generative AI knowledge production. This deficit is combined with philosophical misconceptions about the growth of knowledge using generative AI. This ignorance can give rise to problems in science and society that's why it is imperative to point out the flaws in our existing knowledge.

The potential of generative AI is what makes it even more important to have sufficient awareness of how current Generative AI works, what it can and cannot do. Generative AI works on the ideas of probability to a statement which is in line with an observation. Knowledge grows exactly the other way round. Observation does not precede theory, theory precedes observation (Velthoven & Marcus, 2024). Knowledge grows both in an individual mind as well as in society. AI only engages in calculation without engaging in deeper questioning or explanations. AI cannot advance science as it cannot explain why certain theories should be preferred over others. AI can nevertheless play an important supportive role in science, but AI cannot formulate new scientific explanations (Ordinary Language Philosophy, 2020). Current AI is built on principles on inductivism therein it cannot create new 'universal' or 'probable' statements based on individual instances. Generative AI systems give people the wrong impression that what AI does is similar to or even identical to how humans generate new knowledge.

In the field of generative AI, every individual instance results in an adjustment of the probability distribution. This means that specific instances impact the predictions for future events. The AI model's performance generally depends on the data we provide it with. In contrast, people acquire knowledge by virtue of explanations, not by 'extrapolating' from a collection of thousands of anecdotes. Therefore, AI should be considered as an instrument rather than a creator of new explanatory knowledge. Generative AI models

optimise their training task whenever they create samples that are close to the distribution of the training dataset. This is still entirely different than the way in which humans generate new ideas. When people generate new ideas, their new ideas may be completely unlike any idea that existed before.

The incorporation of generative AI into applications should be built with guardrails in place to address misconceptions and maintain intersectional accountability, recognising that the operation of AI is not akin to human thinking. Treating AI as ''just'' a technology also implies that the ultimate responsibility to explain the applications of AI should remain with people, not with the AI itself. A failure to do so could result in various adverse consequences, such as a wrong use of technology, lack of accountability and transparency, discrimination, and a mismanagement of expectations. Specifically, a misunderstanding of the underlying mechanism of knowledge creation can result in poor regulation and governance strategies (Velthoven & Marcus, 2024).

8. Reflections

In the rapidly evolving landscape of education, language learning is deeply intertwined with culture and identity. For centuries, the teaching and learning of languages has not merely been about acquiring vocabulary or mastering grammar but about immersing oneself in the social, historical, and cultural dimensions of the language in question. Yet, with the recent surge of generative AI, technology that threatens to reduce language learning to a sterile, contextually limited, and ultimately hollow experience. Generative AI, while powerful, should not be included in language learning due to its philosophical, cognitive, and pedagogical limitations, which are rooted in a misunderstanding of the nature of language, culture, and knowledge creation. At its core, language learning is not just about stringing together words or processing linguistic input. It is a dynamic process that serves as a key to cultural understanding and human connection. Language exists within culture, not alongside it, and is a force that can catalyse social change.

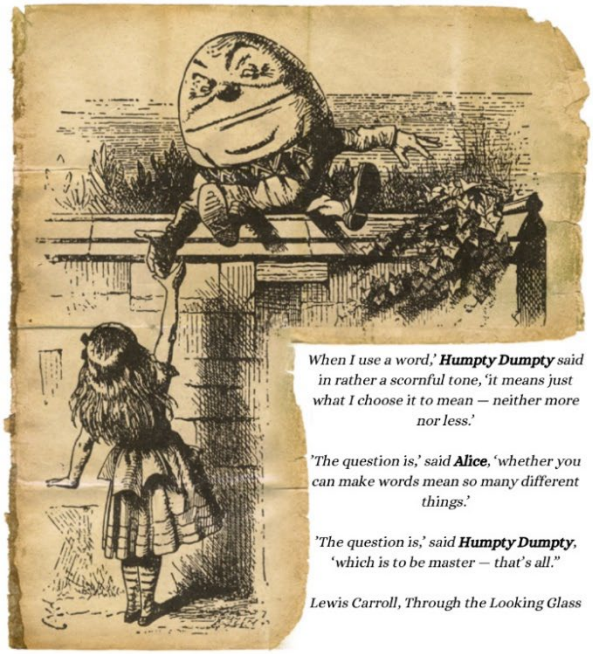
Generative AI, on the other hand, is fundamentally detached from these human experiences. It operates in a vacuum, relying on textual inputs to generate outputs, devoid of the rich, nuanced context that real-life language learning requires. Language learning is not a matter of extrapolating from a dataset of phrases or sentences. It involves constructing mental models of the worlds in which languages operate, shaped by human experiences, emotions, and cultural interactions. AI, however, cannot create new explanatory knowledge; it merely optimises its outputs based on the data it has been trained on. AI cannot think in a language, it only mimics thought and predicting knowledge is not knowledge. This philosophical deficit becomes especially apparent when considering the role of culture in language learning. Language and culture are adaptive mechanisms for human survival and socialisation. Cultural practices, beliefs, and values are transmitted through language, and these elements are not static, they evolve and adapt to new contexts over time. A language learner cannot simply acquire words and syntax; without immersing themselves in the cultural norms and values that those words encode. By reducing language learning to a transactional, probabilistic process, AI strips away the cultural richness that makes language learning a profound and transformative experience (Jenks, 2024).

The introduction of AI into language learning raises ethical concerns regarding transparency, accountability, and equity. AI systems are only as good as the data they are trained on, and if that data is biased or incomplete, the AI's outputs will reflect those deficiencies. This can lead to distorted representations of language and culture, reinforcing stereotypes or excluding certain voices from the language-learning process. Furthermore, the opaque nature of AI decision-making makes it difficult to hold these systems accountable when they fail to meet the needs of learners, potentially leading to a mismanagement of expectations and a lack of trust in the educational process. Moreover, the use of AI in language learning creates a false equivalence between human cognition and machine learning. AI systems give the impression that they can generate new knowledge in the same way humans do, but this is a dangerous misconception.

AI should be viewed as a tool rather than a replacement for human-led language education. It can at the most play a supportive role in certain aspects of language learning, such as vocabulary reinforcement or pronunciation practice, but it cannot replicate the deep, meaningful, and culturally immersive experience of learning a language in its true context. The ultimate responsibility for language education must remain with human educators, who can provide the cultural and contextual knowledge that AI systems lack. By relying on AI for language education, we risk reducing language to a sterile, transactional process that overlooks the deeper cultural and philosophical dimensions of learning.

The analogy drawn from a conversation between Humpty Dumpty and Alice in 'Through the Looking-Glass' highlighting the tension between fixed and fluid meanings of language, which mirrors my discourse analysis on the role of AI in language learning.

Image 5.



In this conversation, **Humpty Dumpty** asserts, "When I use a word, it means just what I choose it to mean—neither more nor less." Highlighting his belief in the arbitrary nature of language, suggesting that words do not have inherent meanings but rather meanings that are assigned by us.

Alice, on the other hand, represents a conventional view of language, wherein words have specific meanings that are universally accepted and understood; assuming an objective reality to the meaning of word.

This is in direct contrast to **Humpty Dumpty**'s subjective and individualistic approach. **Humpty Dumpty**'s perspective is essentially about the relationship between words and their meanings. Instead of words having a fixed meaning and us learning about what they signify; suggesting that the individual should be allowed to dictate what words mean, bending language to personal will and context.

This philosophy challenges the conventional semantics and destabilises the common belief that language is a shared resource with agreed-upon meanings.

Image 6.

Humpty Dumpty presents a nuanced understanding of naming and language exploring the fluidity and malleability of linguistic significance from a distinctly personal perspective.

Humpty Dumpty's perspective on language, represents a form of linguistic relativism, where the meanings of words are not fixed but are subject to individual interpretation.

Alice then contrasts this perspective with a traditional, semantic view that sees language as a set of shared symbols with common meanings that is used for standardised communication among people.

Humpty Dumpty advocates for a radical form of authorial control over language where the speaker has the ultimate authority over the use of words suggesting a personalised somewhat solipsistic relationship with language where intent is in foreground independent of conventional or societal norms.

Alice's is a normative stance where words have stable, consistent meanings that are agreed upon within a linguistic community and is used as a tool for achieving understanding and cooperation between individuals.

Humpty Dumpty's approach challenges the idea of language being a transparent medium for expressing truth or reality. Instead, it suggests that language is a manipulable construct, shaped and reshaped by users.

In postmodernist and constructivist theories about language the role of language is seen as a tool for constructing reality rather than merely describing it. NLP models of AI are built on a **Alice** like perspective where language, words have a fixed meaning that are understood and processed based on statistical regularities in data.



Image 7.



Language is a tool for constructing realities rather than merely describing them. Generative AI does not just interpret input(words) but actively constructs outputs that influence perceptions and realities.

The ability of AI to generate content that may redefine terms or concepts in new and unforeseen ways parallels **Humpty Dumpty**'s notion of linguistic control.

Generative AI models need to be built on **Humpty Dumpty**'s model of language wherein language can be contextualised, personalised and meaning can change based on the speakers intent or the situational context.

Generative AI models need to be built on **Humpty Dumpty**'s model of language wherein language can be contextualised, personalised and meaning can change based on the speakers intent or the situational context.

To be able to do so AI would have to not only determine the literal meaning of the words but also understand the speakers intentions, circumstantial cues that might even alter the literal meaning. This could be done by incorporating user specific data and situational awareness. Embracing the **Humpty Dumpty**-like approach to language would lead to the development of more personalised AI systems that adapt to individual users' language and meanings. These systems would then need to become adaptive and sensitive to the nuances of personal expression and contextual interpretation.

A potential ethical problem could be that because meaning is fluid and controlled by individuals (programmers of AI systems) AI could be misused in shaping public discourse. How the AI models are trained, the data inputted, and the intention of the developer could influence how these systems understand and use language, potentially leading to biases, misunderstandings, or manipulations of information.

Image 8.

Humpty Dumpty's perspective empowers the individual to use language creatively, allowing for personal expression and the capacity to influence and redefine terms within one's own communicative context. It acknowledges the fluid and dynamic nature of language, reflecting how meanings can evolve over time or in different contexts, aligning with the postmodernist understandings of linguistics and semantic change.

Nonetheless there is potential threat in hindering effective communication. If every human being could assign personal meanings to words, mutual understanding could be jeopardised, making communication chaotic and ineffective.

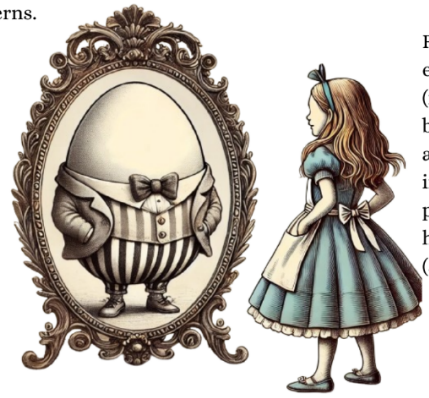
Since it ignores social conventions and shared meanings, **Humpty Dumpty**'s approach could be seen as an extreme form of linguistic relativism. It neglects the communal and social aspects of language that facilitate coordinated and cohesive social interaction. Its practicality is also in question when considering shared concepts such as law, science, and education. There is an arbitrary nature of the linguistic signs which AI does not understand. It does not understand the connection between the signifier (sound pattern) and the signified (concept) that is arbitrary and socially constructed. Mutability and immutability of language signs, which evolves through social use but retains some level of arbitrariness. These signs are not fixed but shaped by context and speaker intent. The idea of language evolving yet constrained by tradition inevitably creates balance between linguistic creativity and the need for stable, shared communication. The system of language is inherently relational—where meanings are defined not by intrinsic value but by their differences from other signs—needing AI to move beyond conventional semantics. AI models, under Saussure's perspective, would then need to understand language as a network of shifting relations rather than a collection of static definitions (Saussure, 1959).

Alice represents a neo colonial, consensus-based view of language, where words have fixed, stable meanings agreed upon by a linguistic community that threatens the existence of non- American cultures and usage of these languages. Her perspective aligns with conventional semantics, seeing language as a shared system that facilitates understanding and cooperation among people.

Image 9.

Alice's perspective ensures clarity and efficiency in communication, relying on shared definitions that are globally understood, reducing misunderstandings and increasing effective communication; hence playing a crucial role in the identity and cohesion of communities. But this could at the same time restrain linguistic evolution, potentially stifling the natural dynamism of language as it evolves with societal changes. It would also exclude or marginalise non-standard dialects and the innovative use of language, which can be important for artistic and personal expression. By strictly adhering to agreed-upon meanings, it limits how individuals can express unique perspectives or adapt language to new contexts and ideas.

A middle ground between **Humpty Dumpty** and **Alice** could be one where we acknowledge that while individuals can influence language, the community ultimately plays a critical role in stabilising meanings through consensus and usage patterns.



Furthermore, modern linguistic theories, proposed by Saussure, emphasises the arbitrary nature of signifiers (words) and their signifieds (meanings). From this perspective, **Humpty Dumpty**'s approach may not be as radical as it initially appears; it simply takes the inherent arbitrariness of language to its logical extreme by personalising it. By incorporating flexibility into AI models, this research pushes for a more personalised approach to language processing, aligning it closer to how humans generate meaning and build relationships through language (Saussure, 1959).

Acknowledgements

This paper is lovingly dedicated to my mother, who sat with me every day as a child, patiently writing down the words I invented. She wrote down each word, along with the meaning I had assigned to them, and then rewrote new ones every day, as I forgot everyday what I had created the day before. She made me believe that my thoughts, my language, and my words held immense power, even at the age of five. Her unwavering faith in my ability to create and the fact that she wrote them all down in a notebook is what has stayed with me till date. Thank you for celebrating each new word and meaning, and for never questioning their place in our conversations. I also want to thank my father for never questioning why his daughter was using hundred new words every day, treating it as the most natural thing in the world. Normalising it is what shaped my unrelenting trust in my ability to create.

References

Beals, J. (2024). Grounding AI: Understanding the implications of generative AI in world language & culture education.

Estes, G. A. (2022). Linguistic flexibility: How a more fluid approach to classroom language can help students succeed.

Jenks, C. J. (2024). Communicating the cultural other: trust and bias in generative AI and large language models.

Language and testimony in classical Indian philosophy. (2020). Stanford Encyclopaedia of Philosophy. [Online]. Available at: https://plato.stanford.edu/entries/language-india/

Namburi, S. & Hopkins, G. (2023). Meraki: Encouraging Language Learning in Real-World Simulations through AI Para-Social Relationship Building

Ordinary Language Philosophy. (2020). Internet Encyclopaedia of Philosophy. [Online]. Available at: https://iep.utm.edu/ord-lang/

O'Regan, J. P. & Ferri, G. (2024). Artificial intelligence and depth ontology: implications for intercultural ethics.

Saussure, F. de. (1959). Course in general linguistics. Edited by C. Bally and A. Sechehaye. Translated by W. Baskin.

Velthoven, M. J. & Marcus, E. J. (2024). Problems in AI, their roots in philosophy, and implications for science and society.

GenAI Acceptance in Professional Services: The Case of Management Consulting

Dennis Nesemeier
Technical University Munich, Germany

dennis.nesemeier@tum.de

Abstract: The increasing relevance of Generative Artificial Intelligence (GenAI) in professional services, particularly in management consulting, and its impact on client services and operations raises the question: *"What factors influence consultants' acceptance of GenAI?"* This study explores the intricate factors influencing the acceptance of GenAI, specifically focusing on management consulting. Using an adapted version of the Unified Theory of Acceptance and Use of Technology (UTAUT) as the theoretical framework, a mixed-methods approach was employed: Twenty semi-structured interviews and a quantitative survey of 140 consultants reveal insights into consultants' perceptions and interactions with GenAI. The findings indicate the relevance of performance and effort expectations, social influence, facilitating conditions, and concerns about GenAI's trustworthiness. Highlighting the complexity of human-technology dynamics, some consultants view GenAI as an opportunity to gain a competitive advantage in their career progression, while others report feeling ashamed when disclosing their use of it. This study broadens the scope of technology acceptance research, introduces specific adaptations and extensions of the theory to better fit the GenAI context, and provides practical managerial recommendations.

Keywords: GenAI, Technology acceptance, Professional services, Management consulting, UTAUT

1. Introduction

The professional service sector and management consulting have undergone a significant transformation with the integration of data-driven analytics in client solutions and operations (Yang et al, 2024). GenAI, often called the next *"digital megatrend,"* is projected to disrupt the consulting landscape significantly (Nissen, 2017, p. 102). A report by the McKinsey Global Institute (2023) suggests that GenAI could lead to an annual productivity gain of 6.1 to 7.9 trillion USD, with between 150 and 250 billion USD specifically within the professional services sector. Dell'Acqua et al (2023) note that using GenAI for specific tasks can significantly enhance the productivity and quality of a consultant's work, enabling them to accomplish 12% more tasks in 25% less time and with a 25% higher quality. Yet, adopting GenAI technologies is a complex endeavor that should not be underestimated: Many related projects have failed to meet their targeted outcomes, often due to complexities related to human-technology dynamics (Vial et al, 2023). Understanding and managing these dynamics is crucial for successful GenAI acceptance in management consulting, leading to the research question: "*What factors influence consultants' acceptance of GenAI?"*

Several scholars have proposed theoretical models to understand factors influencing technology acceptance (e.g., Compeau et al, 1999; Davis et al, 1992; Harrison et al, 1997). In this context, technology acceptance can be understood as a process where individuals' reactions to technology influence their intention to use it (i.e., behavioral intention), which is a predictor of actual use (Davis, 1989). This study uses the UTAUT by Venkatesh et al (2003) as the theoretical framework because it synthesizes key theoretical models while enhancing the explainability and predictability of technology acceptance. The model comprises four main constructs that directly or indirectly impact the behavioral intention to use technology: performance expectancy, effort expectancy, social influence, and facilitating conditions (see Table 1). It is common to extend the UTAUT to fit specific contexts better (see Attuquayefio and Addo, 2014, for an overview).

Table 1: Definitions of the UTAUT's Main Constructs

Main Construct	Definition
Performance Expectancy	*"The degree to which an individual believes that using the system will help him or her to attain gains in job performance."*
Effort Expectancy	*"The degree of ease associated with the use of the system."*
Social Influence	*"The degree to which an individual perceives that important others believe he or she should use the new system."*
Facilitating Conditions	*"The degree to which an individual believes that an organizational and technical infrastructure exists to support the use of the system."*
Note. Definitions taken from Venkatesh et al, 2003.	

2. Method

An exploratory sequential design was chosen for the study as it allows for extending insights from a small group in the initial phase to a broader population in the subsequent phase (Creswell & Clark, 2011). The overall study comprises two parts: (1) The initial qualitative study explored consultants' perceptions and interactions with GenAI through semi-structured interviews. The interview data was then used to tailor the UTAUT to the consulting industry by incorporating relevant aspects. (2) The quantitative study followed, testing and analyzing the previously identified factors for their interplay and relevance to consultants' behavioral intention.

2.1 Qualitative Study Design

The author's professional network was used to identify interviewees, with consultants from seven consultancies agreeing to participate (i.e., Accenture, BCG, EY-Parthenon, McKinsey, Oliver Wyman, Roland Berger, and Strategy&). Twenty interviews were conducted between October and November 2023. Subsequently, the dataset was analyzed following a deductive structured-content analysis. This process resulted in 607 codes distributed among five first-level, sixteen second-level, and four third-level codes (see Figure 1). Social influence was subdivided into two second-level codes (i.e., subjective norm and image) and four third-level codes based on their thematic distinction. Additionally, the analysis led to six control variables, three extending the UTAUT. A new first-level code, trustworthiness, was introduced and focused on confidence in GenAI's reliability, accountability, and ethical conduct. For further validation, the second-level codes of trustworthiness were aligned with the European Commission's Ethics Guidelines for Trustworthy AI (European Commission, 2019).

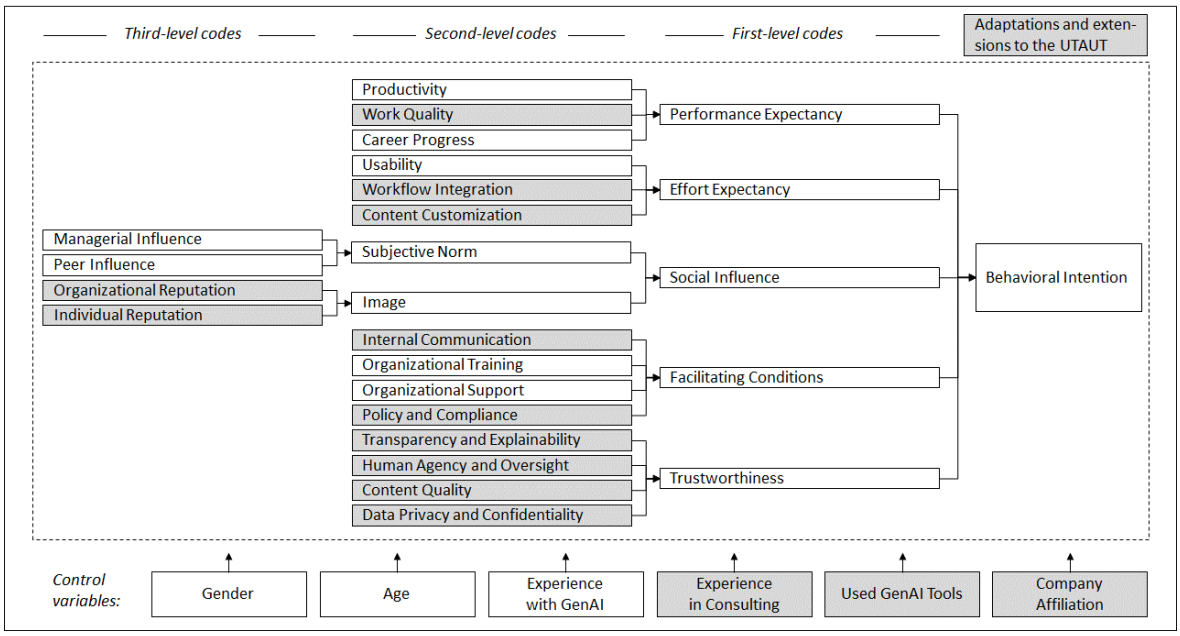


Figure 1: GenAI Acceptance Model

2.2 Quantitative Study Design

With the survey, the factors identified in the qualitative study were tested and analyzed for their interplay and relevance. The survey design follows Venkatesh et al (2003), with adaptations and extensions gleaned from the interviews. The survey was distributed to previous interview participants, requesting internal dissemination to a group that was as representative as possible (i.e., chain referral sampling). From December 2023 to January 2024, a total of N = 140 provided responses. The completion rate was 85%, leading to N = 121 complete responses and a response rate of 65% based on the mentioned contact numbers. The survey comprised 27 questions: six on control variables (see Table 2), 20 on the main constructs (see Tables 3-7), and one on the dependent variable (DV), consultants' behavioral intention (see Table 8). The control variables cover participants' demographics, which reflect the typical tenure track, with a larger base of younger consultants forming the majority while a smaller proportion of senior professionals and partners fall within the higher age range. However, the gender distribution in consulting tends to be more evenly distributed than in the sample.

Table 2: Descriptive Statistics of Control Variables

Gender	**Male**	**Female**	**Diverse**	**Prefer not to say**
	62%	34%	3%	1%
Age	< 20 years	20-30 years	31-40 years	> 40 years
	n/a	**64%**	25%	11%
Experience in Consulting	< 2 years	2-4 years	4-6 years	> 6 years
	37%	**43%**	11%	9%
Company Affiliation	Company A	Company B	Company C	Company D
	35%	13%	19%	11%
	Company E	Company F	Company G	
	9%	5%	9%	
Experience with GenAI	Limited	Basic	Intermediate	Advanced
	6%	**64%**	23%	7%
Used GenAI Tool	Not using GenAI tools	Primarily generic/ public tools (e.g., ChatGPT)	A mix of public and in-house tools	Mainly in-house tools
	6%	**48%**	37%	9%

Note. The most frequent response is highlighted in bold. *N* = 140.

Following Hair et al (2013), a Structural Equation Model (SEM) was chosen to explore multivariate relationships between the main constructs, control variables, and the DV. The survey data was analyzed using Stata 18.0 software. In preparation, the reliability coefficients of the main constructs were assessed to ensure consistency and fit with the research model, and an Exploratory Factor Analysis (EFA) was employed to identify underlying relationships between variables and confirm the model's coherence. As improved reliability coefficients and the EFA suggested, social influence was divided into subjective norm and image for further analysis. Relevant assumptions of the SEM, including linearity, outliers, multicollinearity, homoscedasticity, and normal distribution of the residuals, were also tested. Scatter plots indicated a linear relationship between independent variables and the DV. Cook's Distance (CD = 0.008) confirmed no significant distortions from outliers. Variance-Inflation-Factor (VIF) values under five addressed multicollinearity concerns. Residual-versus-Fitted (RvF) plots revealed heteroscedasticity, which was addressed by using robust standard errors in the model to ensure reliable results. The model was suitable for examining the interplay and relevance of identified factors regarding consultants' behavioral intentions, as tested by the assumptions.

3. Findings

The findings of the qualitative and quantitative studies are described below. First, consultants' general usage of GenAI is outlined. Then, findings are presented according to the model's main constructs: performance expectancy, effort expectancy, social influence (i.e., subjective norm and image), facilitating conditions, and trustworthiness. This is followed by insights on consultants' future expectations and behavioral intention as well as the SEM findings that present a further understanding of multivariate relationships between the main constructs, control variables, and the DV.

3.1 General Usage of GenAI

As indicated by interviews, consultants' usage of GenAI varied from sporadic to daily. Thereby, they use GenAI for various tasks, ranging from information retrieval and content creation to more sophisticated tasks, such as creative thinking, analytical support, and problem structuring. One consultant explained, *"I always see it as a Google alternative, and I think the most important thing here is to save time, get results directly, and save clicks on different pages. So, the bundling of information."* Another consultant mentioned that GenAI provides *"[…] a completely different direction of thoughts."* Furthermore, the survey showed that most consultants had basic (64%) or intermediate (23%) experience with GenAI while mostly using generic/public tools (48%) or a mix of public and in-house tools (37%).

3.2 Performance Expectancy

Described as a *"mini consultant"* or *"co-pilot,"* GenAI was seen positively by many, with 38% strongly agreeing that it is useful in their jobs and 40% strongly agreeing that it increases their productivity. Many consultants noted a positive effect on work quality (33%), though a significant number remained neutral (17%). Additionally, most consultants (41%) viewed GenAI's influence on career prospects as neutral. At the same time, several consultants agreed that it enhances their career prospects (i.e., 19% slightly agreed and 18% agreed). One consultant remarked, *"I believe it gives you an internal advantage to be good at these tools […] there is simply an internal competition. And this is also decided by AI."*

Table 3: Performance Expectancy

	1	2	3	4	5	6	7
I find GenAI useful in my job.	1%	1%	1%	3%	20%	36%	**38%**
Using GenAI increases my productivity.	1%	1%	1%	8%	19%	30%	**40%**
Using GenAI enhances the quality of my work.	1%	1%	5%	17%	**33%**	21%	22%
Using GenAI enhances my career prospects.	1%	4%	7%	**41%**	19%	18%	9%

Note. The most frequent response is highlighted in bold. *N* = 121. The Likert scale was accompanied by the classification: 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neither agree nor disagree, 5 = slightly agree, 6 = agree, 7 = strongly agree.

3.3 Effort Expectancy

Most consultants found GenAI tools user-friendly, with 47% agreeing that they are easy to use. Moreover, 50% found learning to operate them easy. In interviews, many highlighted the importance of GenAI's integration into familiar applications. About 41% agreed that integrating GenAI tools into their existing workflows was easy, but customization of generated content to fit individual requirements was less easy (i.e., 31% slightly agreed and 16% slightly disagreed). Specifically, there were concerns that using generic GenAI content could lead to less precise results, noting that consultants' work *"[...] is often really client-specific, which is difficult for AI."*

Table 4: Effort Expectancy

	1	2	3	4	5	6	7
I find GenAI tools easy to use.	n/a	1%	3%	4%	20%	**47%**	25%
I find learning to operate GenAI tools easy.	n/a	2%	5%	4%	21%	**50%**	18%
I find it easy to integrate GenAI tools into my current workflow.	1%	4%	8%	10%	26%	**41%**	11%
I find it easy to customize generated content to meet my requirements.	1%	2%	16%	15%	28%	**31%**	7%

Note. The most frequent response is highlighted in bold. *N* = 121. The Likert scale was accompanied by the classification: 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neither agree nor disagree, 5 = slightly agree, 6 = agree, 7 = strongly agree.

3.4 Social Influence

As mentioned in section 2.2, social influence was divided into subjective norm and image.

3.4.1 Subjective norm (SN)

Perspectives on managerial and peer influence varied. As indicated by interviews, some managers did not proactively embrace GenAI usage, while others encouraged it through knowledge-sharing sessions. Generally, consultants mentioned the desire for more proactive management endorsement. In the survey, most consultants (37%) were neutral about the positive influence of managers. At the same time, several consultants agreed that managers positively influence their usage of GenAI (i.e., 31% slightly agreed and 17% agreed). Regarding peer influence, 34% agreed it facilitated their usage, with the observation that *"[…] the biggest influencers are those who use it daily."*

3.4.2 Image (IM)

Concerns about GenAI's impact on professional and organizational image led some consultants to hesitate to disclose their use of the technology. In the survey, 31% were slightly worried about the organizational image and 23% about their professional image, as expressed by sentiments such as *"[…] but somehow, you still feel ashamed when you say you used it" or "[…] I should not use the tool live in front of clients […] I would also feel uncomfortable if the window was open.*"

Table 5: Social Influence

	1	2	3	4	5	6	7
SN: My managers positively influence my usage of GenAI.	n/a	2%	3%	**37%**	31%	17%	9%
SN: My peers positively influence my usage of GenAI.	n/a	2%	2%	19%	31%	**34%**	12%
IM: I fear disclosing the use of GenAI could harm my organization's image.	5%	10%	25%	19%	**31%**	8%	2%
IM: I fear disclosing the use of GenAI could harm my professional image.	5%	18%	**31%**	16%	23%	6%	2%
Note. The most frequent response is highlighted in bold. *N* = 121. The Likert scale was accompanied by the classification: 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neither agree nor disagree, 5 = slightly agree, 6 = agree, 7 = strongly agree.							

3.5 Facilitating Conditions

Facilitating conditions for GenAI usage were moderately perceived. Most consultants slightly agreed (37%) or expressed a neutral stance (26%) regarding being well-informed about GenAI, reflecting a reserved view on the adequacy of knowledge dissemination. Additionally, many consultants expressed concerns about getting satisfactory training in GenAI usage from their organizations, with 35% disagreeing that they felt well-trained. Consultants stressed that *"[...] it is crucial to be trained on what is expected of [them] as a skill."* Moreover, slight confidence in the received support from organizations in using GenAI was indicated by 32% slightly agreeing and 27% being neutral about being well-supported. There was also reasonable confidence in complying with existing GenAI policies within their organization, with most consultants slightly agreeing (34%). However, this confidence was contrasted by reservations (i.e., 13% slightly disagreed and 10% disagreed). One consultant confirmed, *"[...] so we have received relatively clear guidance that GenAI cannot actually be used with client data under any circumstances, which we also adhere to by not uploading anything."*

Table 6: Facilitating Conditions

	1	2	3	4	5	6	7
I feel well-informed about GenAI by my organization.	1%	14%	10%	26%	**37%**	12%	n/a
I feel well-trained in using GenAI by my organization.	3%	**35%**	10%	28%	16%	7%	n/a
I feel well-supported in using GenAI by my organization.	2%	16%	14%	27%	**32%**	9%	n/a
I feel confident about complying with my organization's policies on GenAI.	2%	10%	13%	22%	**34%**	19%	n/a
Note. The most frequent response is highlighted in bold. *N* = 121. The Likert scale was accompanied by the classification: 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neither agree nor disagree, 5 = slightly agree, 6 = agree, 7 = strongly agree.							

3.6 Trustworthiness

Concerns emerged regarding the transparency of content creation, with 24% slightly disagreeing and 23% disagreeing that they feel comfortable with the transparency of GenAI's content generation. Moreover, consultants felt slightly positive about the balance between human oversight and GenAI autonomy, with 26% slightly agreeing. One consultant stressed that *"[…] the human must be the ultimate controlling authority.*"

Confidence in the quality and accuracy of generated content was mixed, with most consultants slightly agreeing (31%) in trusting GenAI to generate content of good quality and accuracy, however, a notable number slightly disagreed (21%) and disagreed (10%). Finally, a concern regarding confidence in GenAI to safeguard data and maintain confidentiality was indicated, with more than half of the consultants (68%) either slightly disagreeing (27%), disagreeing (30%), or strongly disagreeing (11%).

Table 7: Trustworthiness

	1	2	3	4	5	6	7
I feel comfortable with the transparency of GenAI's content generation.	5%	23%	**24%**	18%	19%	7%	3%
I feel comfortable with the current balance between human oversight and AI autonomy.	2%	6%	12%	25%	**26%**	23%	7%
I trust GenAI to generate content of good quality and accuracy.	2%	10%	21%	23%	**31%**	11%	2%
I trust GenAI to safeguard data and maintain confidentiality.	11%	**30%**	27%	14%	7%	8%	3%

Note. The most frequent response is highlighted in bold. $N = 121$. The Likert scale was accompanied by the classification: 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neither agree nor disagree, 5 = slightly agree, 6 = agree, 7 = strongly agree.

3.7 Future Expectations and Behavioral Intention

Consultants anticipated significant changes in workflows, expecting GenAI to automate repetitive tasks and allow more time for strategic activities. They foresaw a shift in job roles and a potential workforce reduction. While consultants acknowledged the potential to enhance their work, they emphasized the continuing importance of human talent and customization for client-specific needs. Generally, most consultants intend to use GenAI in the near future, with 34% agreeing and 12% strongly agreeing. One consultant claimed, *"[…] we will no longer be able to turn back this evolution."* Another consultant recalled a quote and stressed, *"[…] if one thinks that this artificial intelligence or OpenAI is just a search engine, then it is as if one had previously thought that a computer is just a calculating machine."*

Table 8: Behavioral Intention

	1	2	3	4	5	6	7
I intend to use GenAI in the near future.	1%	1%	2%	17%	32%	**34%**	12%

Note. The most frequent response is highlighted in bold, and n is displayed in brackets. $N = 121$. The Likert scale was accompanied by the classification: 1 = strongly disagree, 2 = disagree, 3 = slightly disagree, 4 = neither agree nor disagree, 5 = slightly agree, 6 = agree, 7 = strongly agree.

3.8 Structural Equation Model (SEM)

The SEM was used to assess the relative importance of each variable within the GenAI Acceptance Model. As mentioned in section 2.2., relevant model assumptions were tested, as suggested by Hair et al, to ensure the robustness and validity of the model. Given the presence of heteroscedasticity, a robust SEM approach was employed. The SEM confirmed significant positive associations between performance expectancy ($b = 0.25$, $p < 0.01$), effort expectancy ($b = 0.28$, $p < 0.01$), subjective norm ($b = 0.15$, $p < 0.05$), image ($b = -0.17$, $p < 0.05$), facilitating conditions ($b = 0.22$, $p < 0.01$), and trustworthiness ($b = 0.12$, $p < 0.05$) with the DV, consultants' behavioral intention to use GenAI. Moreover, a direct negative effect of age ($b = -0.42$, $p < .001$) and a positive effect of experience with GenAI ($b = 0.36$, $p < .001$) on the DV were observed. Other moderating effects of the control variables, such as those proposed by Venkatesh et al (2003), were not confirmed.

Table 9: Structural Equation Model (SEM)

	Coefficient	Robust Std. Err.	z	P > |z|
Main Constructs:				
Performance Expectancy	0.25***	0.08	3.23	0.001

	Coefficient	Robust Std. Err.	z	P > |z|
Effort Expectancy	0.28***	0.08	3.36	0.001
Subjective Norm	0.15*	0.07	1.98	0.047
Image	-0.17**	0.06	-2.84	0.005
Facilitating Conditions	0.22***	0.07	3.30	0.001
Trustworthiness	0.12*	0.06	2.10	0.036
Control Variables:				
Age	-0.42***	0.11	-3.79	0.000
Gender (Male)	-0.01	0.14	-0.04	0.971
Experience in Consulting	0.00	0.08	0.00	0.997
Experience with GenAI	0.36***	0.099	3.61	0.000
Used GenAI Tool	0.06	0.102	0.56	0.575

Note. Analysis using Stata 18.0. Robust SEM given the presence of heteroscedasticity. Significance level shown as *** = $p < .001$, ** = $p < .01$, and * = $p < .05$.

4. Discussion

The discussion section is divided into theoretical contribution, managerial implications as well as limitations and future fields of research.

4.1 Theoretical Contribution

This study extended the application of the UTAUT by applying it to GenAI within professional services, specifically looking at management consulting. The qualitative study allowed for the transfer of findings from practice to theory while adapting and extending the UTAUT. Aspects mentioned as relevant by interviewed consultants were aggregated through diligent coding into variables used in the quantitative study. These variables included customizing content to fit clients' needs, concerns raised regarding individual and organizational image, and the trustworthiness of GenAI. The latter included aspects such as the transparency of generated content, the balance between human oversight and AI autonomy, the quality of generated content as well as the safeguarding of data and maintaining confidentiality. The quantitative study then facilitated the testing and analysis of these identified factors with a specific focus on their interplay and relevance regarding consultants' behavioral intention to use GenAI. This exploratory sequential design allowed insight from a small group in the initial phase to extend to a broader population in the subsequent phase, thereby creating a comprehensive framework. A framework that other researchers can use in different contexts to further explore and validate findings.

4.2 Managerial Implications

In summarizing key findings, confidence in GenAI's potential to enhance job performance was indicated, alongside some uncertainty regarding its impact on work quality, role profiles, and career prospects. While GenAI tools were generally perceived as user-friendly, there is room for improvement in workflow integration and content customization. Peer influence was perceived as more persuasive than managerial influence, with concerns about GenAI's potential impact on organizational and professional image. There was a foundational level of facilitating conditions, with the desire for enhanced formal training. GenAI's trustworthiness was related to doubts about the transparency of GenAI's content generation process as well as its ability to safeguard data and maintain confidentiality.

These findings can be seamlessly integrated into practical managerial recommendations: Providing practical use cases and best practices for quality assurance and content customization, as well as closely incorporating GenAI into consultants' workflows, can help navigate performance and effort expectations more effectively. Clarifying GenAI's impact on role profiles and career prospects is also recommended to reduce uncertainty and potential reservations about GenAI's impact. Understanding and leveraging social influence is equally important and involves strengthening managerial and peer support, as well as addressing image concerns through open exchange. In this regard, managers should be trained to endorse GenAI, and teams should be encouraged to share their experiences and facilitate mutual learning. Improving facilitating conditions is another key aspect and can be achieved by expanding formal training and providing a dedicated support infrastructure (e.g., peer

support groups, forums, FAQs, or dedicated support teams). Lastly, addressing concerns regarding trustworthiness is essential and can be done through clear communication, comprehensive training, and robust IT governance approaches. However, some technological challenges, such as bias in training data, hallucination, and explainability of outcomes, may not be directly addressable. In such cases, it is crucial to acknowledge these limitations within the implemented measures.

4.3 Limitations and Future Fields of Research

Internal validity was supported by aligning model constructs with established theory and guidelines, and external validity by including a broad set of control variables. However, potential bias related to the sampling method is acknowledged. Therefore, validating findings through representative sampling could improve future research efforts. The findings of this study are based on the consulting industry and may not be generalizable to the entire professional service sector. This limitation suggests applying and validating the model within other sub-sectors, such as legal services, financial services, auditing, or accounting, to enhance external validity.

Moreover, GenAI technology's emergent nature suggests that its long-term impact, especially regarding job roles and skill development, requires further exploration. Specifically, the ease of using technology increases with greater experience, suggesting revisiting the relevance of effort expectations and facilitating conditions over time (Venkatesh et al, 2003).

The construct of trustworthiness was aligned with the European Commission's Ethics Guidelines for Trustworthy AI (2019), which led to the inclusion of certain aspects (i.e., the transparency of generated content, the balance between human oversight and AI autonomy, the quality of generated content as well as the safeguarding of data and maintaining confidentiality). However, due to the exploratory sequential design of the study, some aspects were not explicitly considered as they were not mentioned in the interviews – technical robustness or accountability, to name a few. Future research could explore these dimensions and their impact on technology acceptance further. Given the focus on the trustworthiness of GenAI, analyzing the influential factors of cognitive and emotional trust in the study's context could be further examined.

5. Conclusion

Based on the UTAUT, factors influencing GenAI's acceptance in the professional service sector, specifically in management consulting, were explored. The exploratory sequential design facilitated the identification and analysis of influential factors, contributing to an understanding of consultants' perceptions and interactions with GenAI.

Understanding the nuanced interplay between consultants' perceptions, facilitating conditions, and technology acceptance will enable consultancies to devise more effective strategies for integrating GenAI into their employees' daily routines. This research emphasizes the delicate balance between human-technology dynamics. In this context, not only do corporate actions and managers play a pivotal role in mediating this balance, but consultants do as well. Thus, the research demonstrates that all parties involved are facilitators in effectively adopting and using GenAI tools in consulting practices today.

Looking ahead, the integration of GenAI in management consulting holds immense promise for the industry's evolution. As technology advances and adapts, so will consultants' practices and strategies. This study acknowledges the ongoing development of GenAI applications in management consulting and aims to catalyze further research.

References

Attuquayefio, S., & Addo, H. (2014) "Review of Studies with UTAUT as Conceptual Framework", *European Scientific Journal*, Vol. 10, No. 8, pp. 249–258.

Compeau, D., Higgins, C. A. and Huff, S. (1999) "Social Cognitive Theory and Individual Reactions to Computing Technology: A Longitudinal Study", *MIS Quarterly*, Vol. 23, No. 2, pp. 145–158.

Creswell, J. W. and Clark, V. L. P. (2011) *Designing and Conducting Mixed Methods Research*. SAGE.

Davis, F. D., Bagozzi, R. P. and Warshaw, P. R. (1992) "Extrinsic and Intrinsic Motivation to Use Computers in the Workplace", *Journal of Applied Social Psychology*, Vol. 22, No. 14, pp. 1111–1132.

Davis, F. D. (1989) "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", *MIS Quarterly*, Vol. 13 No. 3, pp. 319–340.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Ra-jendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023) "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality", *Harvard Business School Technology & Operations Management Unit Working Paper*.

European Commission (2019) “Ethics Guidelines for Trustworthy AI”, [online], https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Hair, J. F., Black, W. C., Babin, B. J. and Anderson, R. E. (2013) *Multivariate Data Analysis*, Pearson.

Harrison, D. A., Mykytyn, P. P. and Riemenschneider, C. K. (1997) “Executive Decisions About Adoption of Information Technology in Small Business: Theory and Empirical Tests”, *Information Systems Research*, Vol. 8, No. 2, pp. 171–195.

Nissen, V. (2017) *Digital Transformation of the Consulting Industry: Extending the Traditional Delivery Model*, Springer.

McKinsey Global Institute (2023) “The Economic Potential of Generative AI: The Next Productivity Frontier”, [online], https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier.

Venkatesh, V., Morris, M. G., Davis, G. B. and Davis, F. D. (2003) “User Acceptance of Information Technology: Toward a Unified View”, *MIS Quarterly*, Vol. 27, No. 3, pp. 425–478.

Vial, G., Cameron, A. F., Giannelia, T., & Jiang, J. (2023) “Managing Artificial Intelligence Projects: Key Insights from an AI Consulting Firm”, *Information Systems Journal*, Vol. 33, No. 3, pp. 669–691.

Yang, J., Blount, Y., & Amrollahi, A. (2024) “Artificial Intelligence Adoption in a Professional Service Industry: A Multiple Case Study”, *Technological Forecasting and Social Change*, Vol. 201.

Ecosystem Theory and the Adoption of Artificial Intelligence in SMEs

Steven Nolan and Stelios Zyglidopoulos
Sprott School of Business- Carleton University, Ottawa, Canada

sjvnolan@gmail.com
szyglidopoulos@gmail.com

Abstract: Given the various risks involved in incorporating artificial intelligence (AI) and machine learning into their business operations, firms are at an inflection point about how to do so. In this paper, we propose that for small medium-sized enterprises (SMEs), which lack capital, large quantities of data, and expertise, the best solution would be to join a pre-existing (or developing) ecosystem. From the two potential alternatives available to an SME, going it alone or depending on a larger corporation, we argue for a third option, the joining of an ecosystem of organizations that use AI systems in their operations, as the golden mean. We conclude with some practical and theoretical implications.

Keywords: Artificial intelligence, Adoption of artificial intelligence, SME (small to medium-sized business), Ecosystems, AI research and development, Data autonomy

1. Introduction

Currently, SMEs are making choices about the degree to which they will utilize AI and the ever-expanding capabilities of this new technology (IBM Global AI Adoption Index 2022, 2022; Sjodin, et al 2021). The literature surrounding AI is vast and growing daily. However, when we use an adoption lense we start to see a gap in the literature. Most of the research offers knowledge about large corporations and in particular large tech companies. SMEs make up a very large part of the economy and yet understanding how they are adapting to the introduction of AI has not been examined in as much detail. A significant portion of the literature outlines how ecosystems are used by IT companies for innovation and adopting AI and we became interested to see how ecosystems related to and possibly provided utility to SMEs. Adopting AI technologies is not straightforward, there are many challenges involved and they fall into three key categories. The first challenge is cost. The amount of capital available for R&D has historically been a limiting factor (Schumpeter, 1934). How businesses finance AI adoption (Sharma et al., 2024) and the availability of capital will directly affect the type of AI adopted and its success in providing positive outcomes. The second challenge is the availability of relevant, clean and usable data in sufficient quantities (Jacobides et al., 2021, Hopf et al., 2023). This data is used to train an AI on a particular business function, so it needs to be usable both technically and legally. Finally, one of the biggest challenges in adopting AI is the expertise and knowledge (Alet, 2024) required for a technical system roll out that could have far-reaching effects on business operations. Those who wish to implement AI need to understand both the business problem and how AI can solve it, and such expertise is in high demand (The Human Capital Behind AI, 2021).

Given that SMEs do not usually have access to the above resources, they cannot develop AI alone and must depend on a large tech company for both access to expertise and to significant pools of relevant data. The adoption of AI is surprisingly easy when relying on a large tech firm, however there is a danger in this choice, one that might not be realized for many years, data autonomy is surrendered. Data autonomy in this sense, means a company's ability to use its data exclusively. Large tech firms will pool the data, maybe with the data of other companies within related industries in order to train and make an AI more responsive. In essence, the data that a company generates each year in order to become more competitive is accessible to all through a 'common' AI. Along with 'abdicating the expertise' of how the AI will be trained and what it will focus on to the tech firm, the SME will often have to adopt the large tech firm's platform, perhaps their enterprise software or their cloud computing options, all of these options serve to make it progressively harder for the SME to change services or gain back its autonomy. The speed with which AI is making it into industries and business sectors makes the situation even more precarious to SMEs who may opt for the easiest and fastest choice because of the fear that competitors may already be using AI to gain advantage.

In this paper, we propose a third option, as the golden mean between going it alone and becoming fully dependant on a larger organization for SMEs, which want to implement AI systems in their operations. This third option consists of SMEs joining pre-existing ecosystems of organizations that use AI systems in their operations. Understanding that this recommendation can at first seem counterintuitive, join a group to maintain autonomy, we outline how that can be accomplished.

Ecosystems provide unique opportunities that can create an advantage for firms incorporating AI technologies (Moore, 1993, Williamson & De Meyer, 2012, Müller-Stewens & Stonig, 2019, Holgersson et al., 2022). J.H. Moore defined an ecosystem as "a network of organizations and individuals that co-evolve their capabilities and roles and align their investments so as to create additional value and/or improve efficiency" (Moore, 1993, p.76). Ecosystems are particularly useful to firms when adapting to emerging competitive trends, respond to large technological competitive forces, or developing innovations (Zhang & Williamson, 2021). By joining ecosystems SMEs we argue, can to a certain extent avoid both the need for developing all the necessary resources alone or becoming fully dependent on another larger organization. SMEs gain greater access to experts and have the impetus to develop in-house knowledge, they can pool R&D resources sharing innovations or they can develop parts of a greater whole within the ecosystem. Most importantly they have the ability to retain greater data autonomy.

Ecosystems allow for a deeper inter-relationship between firms, “business ecosystems are an organizational form that enables coordination between independent but complementary players around a shared value proposition, thus creating added value.” (Müller-Stewens and Stonig, 2019, p. 3). There is utility in developing a deeper understanding of business ecosystems to determine if they can be used as a strategy to solve this problem. Questions also arise about how big of an impact the apparent lack of capital for research and development for AI has on businesses, and in particular SMEs. The implications of deepening our knowledge in applying ecosystem theory to the adoption of AI by SMEs are twofold. First, it can offer SMEs new strategy options to help them modernize and remain competitive using AI which given the sheer size of SMEs in the economy, can lead to significant market opportunities. Additionally, a better understanding could guide policymakers and help focus their efforts on supporting SMEs more efficiently and effectively.

The remainder of our paper proceeds as follows. In the next section, we discuss the main problems of AI adoption in business and then we refine our view of these issues to focus on SMEs, specifically the options of exploring AI alone or depending on another larger corporation for its AI adoption. Following, we discuss the main characteristics of ecosystems and proceed to argue that SMEs joining AI ecosystems can avoid, to some extent the previously identified problems. We conclude by discussing the implications of our work for research and managerial practice.

2. Issues Surrounding the Introduction of AI

The business community, government and researchers are all actively discussing AI. According to the IBM Institute, “Roughly half of CEOs feel pressure from business partners to accelerate adoption of generative AI - and two-thirds of executives say their organization needs to quickly adopt generative AI for innovation. However, only 39% of organizations are currently implementing or operating generative AI for innovation and research.” (IBM Global AI Adoption Index, 2022). Despite high interest, adoption lags. Businesses are utilizing this moment to analyze their business needs and identify where AI can add value.

AI availability for business use is extensive with free large language models and bundled options from companies like Microsoft and Adobe. However, adopting commercial off-the-shelf (COTS) AI models may limit overall control and create dependencies on the platform provider. “There are important differences between firms that adopt only off-the-shelf AI solutions and those that build in-house capacity for AI.” (Automation Nation AI Adoption in Canadian Businesses, p. 62). Integrating company data into AI models can complicate future transitions and unintentionally agreeing to a greater relationship with the platform than just the purchase of a service. Data management is crucial (Freeman et al., 2021). To avoid losing competitive advantage, business should consider developing unique, tailored AI strategies for decision-making, process improvement, and innovation. Further, all these COTS options are readily available to an organization's competitors. Therefore, this type of adoption should be entertained as a minimum so that no advantage is surrendered. Simply put, a business could gain a competitive advantage through choosing the harder strategy of developing a more unique and tailored approach, perhaps focusing on how AI can improve its ability to conduct strategic decision making, business process improvement or innovation.

The adoption of AI must be preceded with analysis and understanding the issue. The discrepancy between those business leaders who claim that incorporating AI into their business is very important and those who are moving forward and adopting the new technology can be understood by viewing the problem through the experience lens (The Human Capital Behind AI, 2021). Often an 'issue champion for AI' is needed to shepherd the issue and ensure that the organization is making the right choices in the lead-up to the adoption of AI. There exists the “importance of AI-relevant competencies (e.g., data scientists with deployment-oriented skills or domain experts that can make productive use of data)” (Ångström et al., 2023, p. 7). As the trend of AI adoption increases

so too does the demands on the labour market for those with AI-related skills for example, in Canada it is noted that there has been 17% more hiring in the AI sector than the total labour market from 2018 to 2022 (AI Index Report, 2023).

It has been a long-established norm that firms have difficulty in funding R&D by accessing external funds (Schumpeter, 1934). Carrying debt to conduct R&D presents two distinct issues. First securing debt for R&D requires a higher level of ROI guarantee than for a similar loan intended for capital investment and secondly, the SME needs to demonstrate adequate cash flow to service the debt (Czarnitzki and Hottenrott, 2011).

3. SMEs

By sharing these resources, businesses can significantly reduce the costs associated with acquiring them. Ecosystems can be the “Key to minimizing an important cost disadvantage relative to vertically integrated structures” (Williamson & De Meyer, 2012, p. 33). Adopting the strategy of joining an ecosystem, businesses can share risks and learn from the successes and failures of other ecosystem members (Nalebuff & Brandenburger, 1997). This can reduce costs associated with trial and error (Müller-Stewens & Stonig, 2019) and increase the likelihood of successful innovation and AI adoption. An ecosystem that includes partners, suppliers, and customers helps align those businesses - bringing them all on to the same page “to deliver customer value cost-efficiently in an ecosystem, the activities of partners with complementary capabilities need to be aligned” (Williamson & De Meyer, 2012, p. 35).

Shared costs reduce organizational expenses. Collaboration through an ecosystem can mitigate the largest obstacle to AI adoption: knowledge. Many SMEs lack the understanding to develop, adopt and utilize AI. Approximately 35 to 38 percent of organizations identify the lack of expertise or in-house AI competency as a barrier (Ulrich & Frank, 2021, p. 2156). Conversely, "38% of executives say their organization has the in-house expertise to adopt generative AI for innovation" (Harness the Power of Ecosystem Innovation, IBM Institute for Business Value. https://ibm.co/ceo-generative-ai-innovation).

Most SMEs lack enough AI knowledge needed to assure leadership (Sharma et al., 2024) that expenditures will provide ROI, and that relevant data can be made available for the project. The possibility of proposing changes to a business' core competencies -while the business is operating -will be met with some degree of resistance creating another barrier to the adoption of AI (Alet, 2024). These skill sets are not common (The Human Capital Behind AI, 2021) and many large organizations have been hiring AI experts from academia and tech companies in order to secure their in-house talent. “Undersupply of the relevant skills is known to have first-order effects over innovation and economic performance and shapes the competitiveness of firms, industries and countries alike.” (The Human Capital behind AI, 2021, p. 18). The demand for AI positions in the US labour market tripled over the 5 years from 2014 to 2019 (The Human Capital Behind AI, 2021). Securing in-demand talent is seen to be a greater issue amongst SMEs than for large companies with deep pockets. Access to a cadre of knowledgeable staff will become increasingly important if companies want to remain competitive.

Interestingly, of the three constraints that affect businesses when they consider adopting AI for their organization, the one with the most impact would appear to be access to available, relevant, and clean data (Jacobides et al., 2021, Hopf et al., 2023). Currently, businesses have many options for utilizing AI, mostly commercial off-the-shelf (COTS) solutions. For example, a company can install a customer service chatbot on their IT network, or purchase 'AI as a service' from an external provider, such as an externally hosted LLM or supply-chain advice for a subscription fee. While these options count as AI adoption, they are also available to competitors, offering little competitive advantage. To gain an edge, companies need to develop AI models that are more relevant, unique, and specific to their needs which requires relevant data.

Unfortunately, the lack of this data is a significant hurdle to the development of an AI model (Hopf et al., 2023). The requirement for data that is both relevant to the organization (i.e. customer product preferences) clean (verified and structured in a way to facilitate learning) and exists in large enough quantities makes AI adoption quite difficult for SMEs (Ångström et al., 2023). Data preparation is equally as important as the quantity of data required for machine learning. The data needs to be cleaned, which essentially means it needs to be verified for accuracy and completeness. Lack of data is a common barrier to the adoption of AI as “Organizations are often faced with limited data for machine learning applications. They often need to wait a considerable amount of time, sometimes years, before they have collected sufficiently large and representative amount of data” (Hopf et al., 2023, p. 35). When the data changes or new data is produced, it too needs to be prepared to be useful. Ecosystem data if pooled and applied to the problem of AI adoption, could provide the needed amounts and enhance the AI model and increasing its efficacy (Alet, 2023).

4. Business Ecosystems

Ecosystems provide access to shared resources, perhaps reducing the need to secure outside capital. These resources can include data, technology, infrastructure, and human capital all of which are crucial for innovation and AI adoption. In an environment where resources are finite, collaboration can provide benefits. The structure and modular connectivity of ecosystems lend themselves to collaboration and notably innovation or solving complex problems (Tsvetkova and Gustafsson, 2012). Ecosystems have been documented to provide a competitive advantage to those organizations employing them to solve problems like complex innovation or the adoption of digital platforms. (Adner & Kapoor, 2010, Williamson & De Meyer, 2012).

An ecosystem differs from a value chain (Porter, 1987) and a supply chain (Simchi-Lev, 2005). Value and supply chains involve bilateral, transaction based- relationships. If an organization is removed, the chain can be restored by replacing the organization or internalizing its function. In contrast, removing an entity from an ecosystem has broader implications (Alet, 2023). Additionally, it is important to distinguish between an ecosystem and a digital platform (Jacobides et al., 2021) large companies like Apple and Amazon are often mis-labeled as ecosystems. They are platforms that allow use for a fee or profit-sharing. For example, Apple offers developers code to create applications compatible with iOS profiting from future sales. Apple governs this network (Cusumano and Gawer, 2002) and can remove any developer from it. In its simplest form, "a platform is nothing more than a technological foundation that brings together different activities (of players and objects) in a coordinated way." (Müller-Stewens & Stonig, 2019, p. 5) Due to the requirement of information sharing and coordination - *platforms are necessary for ecosystems, however, platforms aren't necessarily ecosystems*.

The business ecosystem is a force multiplier that provides greater value to the members than could be otherwise achieved through partnerships, supply-chain contracts, or shared digital platforms. The ultimate purpose of an ecosystem is the materialization of a joint value proposition by several players that cannot be achieved by any one of these players in isolation (Jacobides et al., 2018). Organizations can adopt a joint strategy to share resources, knowledge, research, and business data. Understanding of the adoption of AI through ecosystem theory offers solutions to complex problems and also offers practical applications beyond academia.

5. SMEs and AI Ecosystems

There are a lot of conflicting signals about how to adopt AI so a business can gain an advantage. The problems faced by SMEs are capital availability (Sharma et al., 2024), clean and usable data in sufficient quantities (Jacobides et al., 2021, Hopf et al., 2023) and expertise and knowledge (Alet, 2024). Further to the aforementioned common innovation adoption problems, new issues specifically relating to data autonomy, the development of an 'in house' or readily available knowledge cadre, and control over the new AI business tools all make the choices of how a SMEs adopts AI more complex. Business ecosystems, we maintain, can address all these issues better than alternative organizational forms.

Regarding cost, early indication by Fortune 500 companies have shown that AI, when adopted into a business effectively, have reduced costs in “supply chain management (-52%), service operations (-45%), and risk (-43%). On the revenue side, the functions that most respondents saw increases in because of AI adoption were marketing and sales (70%), product and/or service development (70%), and strategy and corporate finance (65%)” (McKinsey -Global AI Survey 2021). The interest in AI adoption is increasing however, cost remains a barrier. Development, training, data-related, infrastructure and maintenance costs all hinder greater adoption despite the benefits and interest. Adopting AI requires a research and development mindset, not just capital investment, significant data preparation is needed. Adopting AI can be a greater commitment than adopting CRM or enterprise software, requiring research, development and project management skills. All of which can provide a disincentive for AI adoption amongst SMEs based solely on the difficulty of this task (Sharma et al., 2024).

AI adoption through a large provider like Microsoft (ChatGPT 4.o) or Amazon (Azure) can take many forms. Developer tools are made available by large tech companies as a way of inducing SMEs to choose one AI over another. However, this does not work for a SME unless they have the requisite knowledge to use those developer tools or a knowledge cadre in their employ. Knowing this, the large tech companies recommend other businesses that, for a fee or by subscribing to an ongoing service, specialize in helping SMEs to adopt a large tech's AI. In both of these arrangements the SME would still be required to surrender their data (past and future) to the large tech firm. This data would of course be pooled and used to enhance the AI and its future training requirements. The problem with this model is twofold. First, data autonomy is surrendered, and full control is lost. Ironically, a SMEs data could conceivably be pooled with their competitors' data using the same AI. This instance would see

a very limited advantage gained by using AI and in fact a loss of profitability because of the ongoing outlay of capital in using the big tech's AI platform.

6. Implications and Conclusion

The integration of AI into business operations has become a strategic imperative. SMEs leveraging ecosystems for AI development can significantly enhance their business processes, leading to a multitude of positive implications. Initially, access to resources and expertise is a key advantage. AI development requires substantial resources and specialized knowledge, which businesses may lack. By participating in an ecosystem, they can access shared resources and tap into the collective expertise of the ecosystem, thereby overcoming resource constraints and accelerating AI development. The interconnectedness and diversity of ecosystems can provide a buffer against uncertainties and changes in the business environment. This can enable firms to adapt and respond effectively to challenges and disruptions in their AI development efforts. It provides access to resources and expertise, fosters innovation, enhances learning and capability development, provides market access and opportunities, and retains much more autonomy for the SME.

Although the use of an ecosystem could lower costs and provide necessary access to data and AI knowledge there are still limitations to what this strategy can achieve. While large corporations have been at the forefront of adopting AI, SMEs have found adoption difficult. SMEs may be reluctant to share their data within the ecosystem due to privacy and security concerns. Choosing to join an existing ecosystem over a large tech firm's platform would give the SME much greater control over the data and how it is handled. A similar critique of joining an ecosystem arises when the future costs of development are considered. Would it not be better to join a large tech firm and allow them to focus on innovation? Unfortunately, once committed to a particular large tech AI the cost can only increase over time as the difficulty a SME would face to leave that platform and move to another would prove to be much greater than staying. A SME may very well take the decision to let large tech 'take care' of their AI, which would have them surrender their data autonomy, make them in some ways 'beholden' to big tech and would not allow them the opportunity to innovate independently of any of their competitors who are also part of the same or similar large tech corporation's AI platform.

The practical implications offered by the approach of joining an ecosystem offer new options to SMEs that are concerned about data autonomy, the development of their own knowledge cadre and maintaining more control over their AI tools. As well, this approach may be of interest to government policymakers who seek to support local or home country SMEs and their ability to maintain their autonomy and competitiveness. More exploration of ecosystem theory, specifically determining the line between what makes an ecosystem and what becomes a platform dominated by one large member could provide great insight to academia on how AI is being adopted both regionally and in different industries. Future studies could explore not only how entrepreneurs are utilizing this powerful business tool but also if AI is helping to create more entrepreneurs or if the only winners of the new AI landscape are the existing large tech firms.

In conclusion, ecosystems, with their collaborative, innovative, and resilient nature, could provide a powerful approach to help solve complex business problems such as the adoption of AI by SMEs. As businesses continue to grapple with increasing complexities in the global business environment, the role of business ecosystems is likely to become even more significant. Seeking to overcome the barriers to AI adoption, organizations may use joining an existing business ecosystem as a strategy to reduce capital costs, share knowledge and perhaps most importantly retain their data autonomy.

References

Adner, R., & Kapoor, R. *(2010*). 'Value creation in innovation ecosystems: How the structure of technological interdependence affects firm performance in new technology generations.' *Strategic Management Journal,* 31(3), 306–333. https://doi.org/10.1002/smj.821

AI Index Report 2023 – Artificial Intelligence Index (2023). Available at: https://aiindex.stanford.edu/report/ (Accessed: 12 April 2024).

Alet, J. (2023) 'Key pillars in a business ecosystem success', *Journal of Business Strategy*, 44(5), pp. 294–300. Available at: https://doi.org/10.1108/JBS-05-2022-0077.

Alet, J. (2024) 'Effective integration of artificial intelligence: key axes for business strategy', *Journal of Business Strategy*, 45(2), pp. 107–114. Available at: https://doi.org/10.1108/JBS-01-2023-0005.

Ångström, R.C. *et al.* (2023) 'Getting AI Implementation Right: Insights from a Global Survey', *California Management Review*, 66(1), pp. 5–22. Available at: https://doi.org/10.1177/00081256231190430.

Automation Nation? AI Adoption in Canadian Businesses (2023) *The Dais*. Available at: https://dais.ca/reports/automation-nation-ai-adoption-in-canadian-businesses/ (Accessed: 4 April 2024).

'Automation Nation? AI Adoption in Canadian Businesses' (no date).
Birkinshaw, J. (2023) 'How Incumbent Firms Respond to Emerging Technologies: Comparing Supply -Side and Demand -Side Effects', *California Management Review*, 66(1), pp. 48–71. Available at: https://doi.org/10.1177/00081256231199263.
Brynjolfsson, E. (2022) 'The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence', *Daedalus*, 151(2), pp. 272–287. Available at: https://www.jstor.org/stable/48662041 (Accessed: 1 April 2024).
Cusumano, M.A. and Gawer, A. (2002) 'The Elements of Platform Leadership: MIT Sloan Management Review', *MIT Sloan Management Review*, 43(3), pp. 51–58. Available at: https://proxy.library.carleton.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=6553425&site=ehost-live (Accessed: 9 May 2024).
Czarnitzki, D. and Hottenrott, H. (2011) 'R&D investment and financing constraints of small and medium-sized firms', *Small Business Economics*, 36(1), pp. 65–83. Available at: https://www.jstor.org/stable/41486077 (Accessed: 12 April 2024).
Elish, M.C. and Boyd, Danah (2018) 'Situating methods in the magic of Big Data and AI', *Communication Monographs*, 85(1), pp. 57–80. Available at: https://doi.org/10.1080/03637751.2017.1375130.
Freeman, L., Rahman, A. and Batarseh, F.A. (2021) 'Enabling Artificial Intelligence Adoption through Assurance', *Social Sciences*, 10(9), p. 322. Available at: https://doi.org/10.3390/socsci10090322.
Hirsch, P.B. (2023) 'At the crossroads: generative AI and corporate risk management', *Journal of Business Strategy*, 44(6), pp. 426–429. Available at: https://doi.org/10.1108/JBS-09-2023-0196.
Holgersson, M. *et al.* (2022) 'The Forces of Ecosystem Evolution', *California Management Review*, 64(3), pp. 5–23. Available at: https://doi.org/10.1177/00081256221086038.
'IBM Global AI Adoption Index 2022' (2022).
Jacobides, Michael G. "How to Compete When Industries Digitize and Collide: An Ecosystem Development Framework." *California Management Review*, vol. 64, no. 3, 2022, pp. 99–123, https://doi.org/10.1177/00081256221083352.
Jacobides, M.G., Brusoni, S. and Candelon, F. (2021) 'The Evolutionary Dynamics of the Artificial Intelligence Ecosystem', *Strategy Science*, 6(4), pp. 412–435. Available at: https://doi.org/10.1287/stsc.2021.0148.
Jacobides, M.G., Cennamo, C. and Gawer, A. (2018) 'Towards a theory of ecosystems', *Strategic Management Journal*, 39(8), pp. 2255–2276. Available at: https://doi.org/10.1002/smj.2904.
Journal of Governance and Regulation (no date). Available at: https://virtusinterpress.org/-Journal-of-Governance-and-Regulation-15- (Accessed: 8 March 2024).
Mantri, A. and Mishra, R. (2023) 'Empowering small businesses with the force of big data analytics and AI: A technological integration for enhanced business management', *The Journal of High Technology Management Research*, 34(2). Available at: https://doi.org/10.1016/j.hitech.2023.100476.
Meyer, A.D. and Williamson, P.J. (2023) 'Kick-starting an Innovation Ecosystem', 03(3).
Moore, J.F. (1993) 'Predators and Prey: A New Ecology of Competition', *Harvard Business Review*, 71(3), pp. 75–86. Available at: https://proxy.library.carleton.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9305180275&site=ehost-live (Accessed: 10 May 2024).
Müller-Stewens, G. and Stonig, J. (2019) 'Business Ecosystems and Platforms: Towards a Shared Understanding', *Die Unternehmung*, 73(4), pp. 381–386. Available at: https://www.jstor.org/stable/26995452 (Accessed: 31 March 2024).
Nalebuff, B. J., & Brandenburger, A. M. (1997). 'Co-opetition: Competitive and cooperative business strategies for the digital economy', *Strategy & Leadership*, 25(6), 28–33. https://doi.org/10.1108/eb054655
'Platform Leadership: How Intel, Microsoft, and Cisco Drive Industry Innovation' (2003) *Innovation*, 5(1), pp. 91–94. Available at: https://doi.org/10.5172/impp.2003.5.1.91.
Porter, M.E. (1987) 'From competitive advantage to corporate strategy', *Harvard Business Review*, 65(3), pp. 43–43. Available at: https://proxy.library.carleton.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=4126877&site=ehost-live (Accessed: 9 February 2024).
Porter, M.E. (1996) 'What Is Strategy?', *Harvard Business Review*, 74(6), pp. 61–78. Available at: https://proxy.library.carleton.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=9611187954&site=ehost-live (Accessed: 10 January 2024).
Schumpeter, J.A. (2021) *The Theory of Economic Development*. 1st edn. Oxford: Routledge (Routledge Classics). Available at: https://doi.org/10.4324/9781003146766.
Sharma, S. *et al.* (2024) 'Why Do SMEs Adopt Artificial Intelligence-Based Chatbots?', *IEEE Transactions on Engineering Management*, 71, pp. 1773–1786. Available at: https://doi.org/10.1109/TEM.2022.3203469.
Sjödin, D. *et al.* (2020) 'Value Creation and Value Capture Alignment in Business Model Innovation: A Process View on Outcome-Based Business Models: Journal of Product Innovation Management', *Journal of Product Innovation Management*, 37(2), pp. 158–183. Available at: https://doi.org/10.1111/jpim.12516.
Sjödin, D. *et al.* (2021) 'How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops', *Journal of Business Research*, 134, pp. 574–587. Available at: https://doi.org/10.1016/j.jbusres.2021.05.009.
The human capital behind AI: Jobs and skills demand from online job postings (2021). OECD Science, Technology and Industry Policy Papers 120. Available at: https://doi.org/10.1787/2e278150-en.

The who, what, and where of AI adoption in America | MIT Sloan (2024). Available at: https://mitsloan.mit.edu/ideas-made-to-matter/who-what-and-where-ai-adoption-america (Accessed: 2 April 2024).

Tsvetkova, A. and Gustafsson, M., 2012. Business models for industrial ecosystems: a modular approach. *Journal of Cleaner Production*, *29*, pp.246-254.

Ulrich, P. and Frank, V. (2021) 'Relevance and Adoption of AI technologies in German SMEs – Results from Survey-Based Research', *Procedia Computer Science*, 192(Complete), pp. 2152–2159. Available at: https://doi.org/10.1016/j.procs.2021.08.228.

Wang, P. (2019) 'On Defining Artificial Intelligence', *Journal of Artificial General Intelligence*, 10(2), pp. 1–37. Available at: https://sciendo.com/article/10.2478/jagi-2019-0002 (Accessed: 1 April 2024).

Williamson, P.J. and De Meyer, A. (2012) 'Ecosystem Advantage: How to Successfully Harness the Power of Partners', *California Management Review*, 55(1), pp. 24–46. Available at: https://doi.org/10.1525/cmr.2012.55.1.24.

Zhang, M.Y. and Williamson, P. (2021) 'The emergence of multiplatform ecosystems: insights from China's mobile payments system in overcoming bottlenecks to reach the mass market', *Technological Forecasting and Social Change*, 173, p. 121128. Available at: https://doi.org/10.1016/j.techfore.2021.121128.

Security and Safety Concerns in the age of AI

Victoria Yousra Ourzik
University of Paris Dauphine PSL, France

victoria.ourzik@dauphine.eu

Abstract: Artificial Intelligence (AI) is transforming industries at an astonishing rate, reshaping how we live, work, and interact with technology. Yet, as AI becomes more pervasive, it brings urgent questions about security and safety. This article explores these critical issues, drawing a clear distinction between AI security and AI safety—two concepts that are often misunderstood but are crucial for responsible AI deployment. AI security focuses on protecting systems from external threats like data breaches, adversarial attacks, and unauthorized access. As AI systems increasingly handle sensitive data and control critical operations, securing them against such risks is essential. A breach or failure could compromise not only privacy but also the integrity of critical infrastructures. On the other hand, AI safety extends beyond technical defenses to the broader societal implications of AI. Issues like algorithmic bias, ethical decision-making, and unintended consequences of AI systems highlight the risks to human well-being. As AI becomes more autonomous, its alignment with human values and societal norms becomes paramount. Furthermore, the existential risks posed by advanced AI—such as loss of control or unintended outcomes—raise profound questions about the future of human-AI coexistence. This article delves into real-world case studies of AI failures and near-misses, offering tangible insights into the potential consequences of unchecked AI growth. It also explores strategies for mitigating these risks, balancing the pursuit of innovation with the need for transparency, accountability, and ethical oversight. As we look to the future, international cooperation and robust regulatory frameworks are essential to managing AI's growing influence. By examining both technical and ethical dimensions, this article equips readers with a comprehensive understanding of AI security and safety, urging a proactive approach to managing the risks and harnessing the potential of this powerful technology.

Key words: Artificial intelligence (AI), AI security, AI safety, Data breaches, Responsible AI

1. Introduction

As technology advances at an unprecedented pace, Artificial Intelligence (AI) has emerged as a transformative force with significant implications across various sectors. While AI offers remarkable opportunities, its integration into everyday life also raises critical concerns about the security and safety of AI systems. These concerns, if not effectively addressed, can overshadow the substantial benefits that AI has to offer. This article aims to provide a comprehensive understanding of AI security and safety, offering practical guidance on how to navigate these complex challenges.

At the core of this discussion are two fundamental concepts: AI security and AI safety. Though these terms are often used interchangeably, they address distinct aspects of AI deployment. AI security focuses on protecting AI systems from malicious exploitation. This includes safeguarding against data breaches, adversarial attacks, and unauthorized access (Goodfellow, Shlens and Szegedy, 2014). Ensuring that AI systems are resilient to these threats is crucial for maintaining the confidentiality, integrity, and availability of sensitive information and critical operations.

In contrast, AI safety extends beyond the realm of cybersecurity to encompass ethical, social, and existential concerns associated with AI technologies. This dimension addresses the impact of AI on human well-being and societal norms. Key issues include algorithmic bias, unintended consequences of AI decision-making, and the alignment of AI operations with ethical principles. Additionally, the potential existential risks posed by advanced AI systems prompt important questions about the long-term trajectory of AI development and its implications for humanity (Floridi et al., 2018).

To provide a thorough exploration of these topics, this article is structured to cover several critical areas. We begin with the categorization of AI threats, examining various types of risks including data breaches, system vulnerabilities, and adversarial attacks, and discuss strategies for their mitigation. Following this, we delve into high-profile case studies that highlight real-world examples of AI-related incidents, offering insights into how these events unfolded and what can be learned from them. Finally, we explore future directions and challenges, focusing on how to advance toward responsible AI, balance performance with transparency, and harmonize regulations across jurisdictions.

By exploring these areas, this article aims to equip readers with the knowledge needed to address the multifaceted dimensions of AI security and safety and to navigate the complexities that arise as AI technology continues to evolve.

2. Categorization of AI Threats

In the relentless progression toward an AI-driven future, it is crucial to understand the broad spectrum of threats facing these systems. This section offers a thorough examination of AI threats, dividing them into three key categories: data breaches, system vulnerabilities, and adversarial attacks. Each category poses distinct challenges and requires specific mitigation strategies to protect AI systems from exploitation and manipulation. By analyzing these threats in detail, we aim to provide stakeholders with the insights necessary to strengthen AI ecosystems against emerging risks and uphold the integrity and reliability of AI technologies.

2.1 Data Breaches

Data breaches represent a critical threat to the integrity and security of AI systems. It is essential for organizations to establish appropriate technical and organizational measures to prevent potential data breaches. However, the reality remains that data breaches are always a possibility. Specifically, a data breach occurs when a company's entrusted data experiences a security incident, resulting in breaches of confidentiality, availability, or integrity. In such cases, there is a significant likelihood that the breach jeopardizes individuals' rights and freedoms. As mentioned earlier, data breaches may encompass financial information such as bank details or credit cards, personally identifiable information (PII), personal health information, as well as corporate trade secrets or intellectual property. Typically, data breaches involve compromised files, documents, and other sensitive information.

2.1.1 Impact

Numerous instances globally, particularly in the United States and Europe, underscore the susceptibility of companies to data security breaches. In 2018, Marriott International, a global hospitality company, disclosed a significant data breach that compromised the personal information of approximately 500 million guests. The breach stemmed from unauthorized access to the Starwood guest reservation database, a subsidiary of Marriott International (Marriott International, 2018). Attackers gained illicit entry into the database, exposing a plethora of sensitive details, including guests' names, addresses, email addresses, phone numbers, passport numbers, and in some cases, payment card information (Krebs, 2018).

The repercussions of the Marriott International data breach reverberated across multiple dimensions, eliciting legal, regulatory, and reputational fallout. In the aftermath of the breach, Marriott faced intense scrutiny from regulatory authorities and the public, prompting investigations into the company's security practices and compliance with data protection regulations. The breach incurred substantial financial liabilities for Marriott, including a hefty fine imposed by regulatory bodies such as the Information Commissioner's Office (ICO) under the General Data Protection Regulation (GDPR) (Herbert Smith Freehills, 2020). Moreover, the incident eroded consumer trust in Marriott's ability to safeguard their personal information, leading to a decline in customer confidence and brand reputation. In addition to Marriott, other notable victims of cyber threats include British Airways (BA) and Equifax, both experiencing large-scale breaches affecting millions of customers. Expedia subsidiary Orbitz also reported a security breach impacting approximately 880,000 payment cards, potentially exposing customers' information to hackers (Thomaidis, 2022). These incidents highlight numerous cases of cybersecurity breaches in prominent hotels and companies within the tourism sector. The fallout from the breach serves as a stark reminder of the imperative for organizations to prioritize robust data protection measures, implement stringent security protocols, and conduct regular security audits to preempt and mitigate the risks of data breaches. Additionally, it underscored the urgent need for enhanced collaboration between industry stakeholders, regulatory bodies, and cybersecurity experts to fortify defenses against evolving cyber threats and uphold the integrity and privacy of individuals' data in an increasingly interconnected world.

2.1.2 Mitigation

In the digital era, data breaches represent a substantial threat to organizations of all sizes, potentially resulting in severe financial losses and reputational damage. Addressing these risks necessitates the implementation of robust mitigation strategies to protect sensitive information. This section explores key approaches to mitigating data breaches, offering a thorough analysis of best practices and technologies aimed at strengthening data security.

Encryption

Encryption serves as a foundational safeguard against data breaches by rendering data unreadable to unauthorized entities. It is imperative to employ robust encryption protocols to protect sensitive data both at rest and in transit. Leveraging strong encryption algorithms and implementing sound key management practices are essential elements of an effective encryption strategy.

Access controls

Implementing stringent access controls is paramount to restrict data access to authorized personnel exclusively. Utilizing mechanisms such as role-based access control (RBAC) and attribute-based access control (ABAC) can help enforce granular access policies. Additionally, incorporating multi-factor authentication (MFA) adds an extra layer of security by requiring multiple forms of verification before granting access.

Regular audits and monitoring

Continuous monitoring and proactive detection of anomalous activities are crucial for thwarting data breaches. Conducting regular security audits enables organizations to identify vulnerabilities and remediate them promptly. Deploying robust intrusion detection systems (IDS) and intrusion prevention systems (IPS) facilitates real-time monitoring of network traffic, allowing for swift response to potential security incidents.

By adopting a multi-faceted approach encompassing encryption, access controls, and continuous monitoring, organizations can bolster their defences against data breaches and safeguard the integrity of their AI systems and the sensitive data they handle.

2.2 System Vulnerabilities

System vulnerabilities refer to weaknesses or flaws in software or hardware that can be exploited by attackers to gain unauthorized access or cause harm. AI systems are particularly susceptible to vulnerabilities due to their complexity and reliance on extensive software components. These vulnerabilities can arise from various sources, including coding errors, inadequate security measures, and misconfigurations. Furthermore, the integration of third-party libraries, open-source tools, and the dependence on vast amounts of data can introduce additional points of failure. Adversaries may exploit these weaknesses to manipulate AI models, execute adversarial attacks, steal sensitive information, or disrupt operations. Effective identification, assessment, and mitigation of system vulnerabilities are crucial to maintaining the integrity and security of AI systems (Goodfellow, Shlens and Szegedy, 2014).

2.2.1 Impact

Exploiting system vulnerabilities can disrupt the normal functioning of AI systems, causing downtime, reduced efficiency, and operational delays, which translate to lost revenue and customer dissatisfaction. Attackers can manipulate AI models by introducing malicious data or leveraging weaknesses in the model architecture, leading to incorrect predictions or decisions. This undermines the reliability of AI applications in critical areas such as healthcare, finance, and autonomous systems. Public disclosure of system vulnerabilities can significantly damage an organization's reputation, resulting in long-term adverse effects on the brand and market position. Organizations may also face stringent penalties and legal actions if vulnerabilities lead to breaches that violate data protection regulations. In critical sectors such as defense and national security, AI system vulnerabilities can be exploited by state actors or terrorists, posing significant threats to national security and public safety.

A real-world example of the impact of system vulnerabilities can be seen in the 2020 Twitter hack. In this incident, attackers exploited vulnerabilities in Twitter's internal systems to gain access to the accounts of several high-profile users, including Elon Musk, Bill Gates, and Barack Obama. The attackers used these accounts to post a cryptocurrency scam, asking followers to send Bitcoin to a specific address with the promise of doubling their money. The breach not only resulted in financial losses for those who fell victim to the scam but also raised significant concerns about the security of social media platforms. Twitter faced intense scrutiny from both the public and regulatory bodies, highlighting the importance of robust security measures and the need to protect sensitive user information from exploitation.

2.2.2 Mitigation

To effectively protect AI systems from vulnerabilities, it is crucial to adopt a comprehensive set of mitigation strategies. This section highlights essential practices to enhance the security and resilience of AI systems.

Regular patching

Ensure timely application of security patches and updates to all software and hardware components. Maintain an inventory of assets to track and manage vulnerabilities.

Security by design

Incorporate security considerations into the design and development of AI systems. Conduct thorough security testing and code reviews to identify and address potential vulnerabilities.

Isolation and segmentation

Implement network segmentation and isolation to limit the impact of a potential breach. Use virtual machines and containerization to create isolated environments for different components.

2.3 Adversarial Attacks

Adversarial attacks involve deliberately manipulating input data to deceive AI models, causing them to make incorrect predictions or classifications. Unlike system vulnerabilities, which pertain to flaws or weaknesses in the software or hardware infrastructure that can be exploited for unauthorized access or damage, adversarial attacks specifically target the algorithms and data processing components of AI systems (Goodfellow, Shlens and Szegedy, 2014). These attacks exploit the inherent susceptibilities of AI models, such as their sensitivity to small perturbations in input data, posing significant threats to the integrity, reliability, and trustworthiness of AI applications.

2.3.1 Impact

Adversarial attacks on AI systems can profoundly compromise their accuracy and reliability, leading to significant and potentially dangerous consequences. These attacks exploit the vulnerabilities inherent in AI models, resulting in erroneous predictions and classifications. The repercussions are especially severe in critical sectors such as healthcare, finance, and autonomous systems, where inaccuracies can endanger lives, cause substantial financial losses, and erode trust in AI technologies. Additionally, adversarial attacks can damage an organization's reputation, necessitating considerable resources for ongoing monitoring and the development of robust defensive measures.

A notable example of the impact of adversarial attacks is the research by Szegedy et al. (2014), which demonstrated that minor perturbations to images could lead AI models to misclassify them. For instance, they showed that an image of a panda could be misclassified as a gibbon by adding almost imperceptible noise. This finding highlights the susceptibility of AI models to slight manipulations and raises significant concerns about their robustness in practical applications.

Another critical example is the study by Grosse et al. (2017), which focused on AI-based malware detection systems. Their research revealed that attackers could alter malware samples in ways that allow them to evade detection by AI systems while preserving their malicious functionality. This ability to bypass security measures without reducing the threat posed by the malware underscores a significant risk to cybersecurity systems that depend on AI.

2.3.2 Mitigation

Mitigating the risks posed by adversarial attacks requires a multifaceted approach to enhance the resilience and robustness of AI systems. The following strategies outline key measures to protect AI models from adversarial manipulations:

Training

Train AI models using adversarial examples to improve their robustness. Techniques such as adversarial training and defensive distillation can enhance model resilience to adversarial attacks.

Detection & response

Develop mechanisms to detect and respond to adversarial attacks in real-time. Implement anomaly detection systems to identify suspicious patterns indicative of adversarial manipulation.

Regular model evaluation

Continuously evaluate AI models against a diverse set of adversarial scenarios. Regularly update models to address new attack vectors and maintain their robustness.

Understanding and mitigating AI threats is crucial for the safe and ethical deployment of AI technologies. Data breaches, system vulnerabilities, and adversarial attacks represent significant challenges that require comprehensive security measures and continuous vigilance. By implementing robust security practices, organizations can protect AI systems and the sensitive data they handle, ensuring their integrity and reliability.

3. High-Profile Case Studies

Analysing real-world AI security failures offers critical insights into the vulnerabilities and risks inherent in AI systems. This section delves into prominent case studies, including the Facebook and Cambridge Analytica data scandal (2018), the fatal accident involving Uber's self-driving car (2018), and the Microsoft Tay Twitter bot incident (2016). By examining these cases, we can better understand the implications of such failures and glean valuable lessons for enhancing AI security practices.

3.1 The Facebook and Cambridge Analytica Data Scandal (2018)

In March 2018, a whistleblower disclosed that Cambridge Analytica, a political consulting firm, had unlawfully accessed the personal data of millions of Facebook users. This breach highlighted critical deficiencies in data security and privacy practices, provoking significant public backlash and regulatory investigation.

3.1.1 How the breach occurred

The origins of the scandal date back to 2014, when Dr. Aleksandr Kogan, a researcher at Cambridge University, developed a personality quiz application named "This Is Your Digital Life." This app, hosted on Facebook, not only collected data from users who interacted with it but also from their friends, exploiting the Facebook API available at that time. While approximately 270,000 users participated in the quiz, the app was able to harvest data from approximately 87 million Facebook profiles due to the API's design (Isaak & Hanna, 2018). Cambridge Analytica subsequently used this extensive data set to construct psychological profiles and deliver targeted political advertisements, allegedly impacting significant political events such as the 2016 U.S. presidential election and the Brexit referendum (Confessore, 2018).

3.1.2 Key takeaways

The Facebook and Cambridge Analytica scandal reveals significant flaws in data protection and regulatory oversight. Facebook's inadequate data protection policies allowed extensive third-party access to user information, facilitating the massive data extraction by Cambridge Analytica. Users were inadequately informed about data collection practices, undermining their privacy rights (Isaak & Hanna, 2018). The absence of clear regulations in the data usage and AI analytics space enabled the breach to go unchecked. Vulnerabilities in the AI systems used highlight the urgent need for stronger ethical guidelines and security frameworks.

The scandal results in severe reputational damage for Facebook, a $5 billion FTC fine, and accelerates global regulatory changes, such as the implementation of GDPR (General Data Protection Regulation, 2018). It also prompts other tech companies to enhance their data security practices. Key lessons include the necessity of robust data governance, transparency, and user control over data. The incident underscores the importance of addressing data protection vulnerabilities and establishing rigorous ethical and security standards to prevent similar breaches in the future.

3.2 The Fatal Accident Involving a Self-Driving Uber (2018)

On March 18, 2018, an autonomous Uber SUV struck and killed Elaine Herzberg, a pedestrian, in Tempe, Arizona. This incident was the first recorded fatality involving a self-driving vehicle and raised significant concerns about the safety and maturity of autonomous vehicle technology for public deployment.

3.2.1 How the incident occurred

The Uber vehicle involved was a Volvo XC90 operating in autonomous mode, with a human safety driver present behind the wheel. The vehicle was traveling at approximately 40 mph when it collided with Herzberg, who was walking her bicycle across the road outside of a crosswalk. Although the onboard sensors detected Herzberg, the vehicle's software misclassified her presence as a false positive, leading to a failure in taking appropriate action to prevent the collision.

The accident highlighted several critical issues with the self-driving technology. The vehicle's sensors and software failed to accurately classify Herzberg as a pedestrian in time to prevent the collision, exposing limitations in the object detection and classification algorithms. The emergency braking system was deactivated to avoid erratic behavior, leaving reliance on the human safety driver, who was distracted at the time, rendering the safety measure ineffective. Moreover, there was inadequate oversight and training for the safety driver, and no system ensured the driver's attentiveness. Additionally, the self-driving system design lacked comprehensive solutions for ethical concerns, such as pedestrian safety and robust fail-safes.

3.2.2 Key takeaways

The 2018 self-driving Uber vehicle accident has profound repercussions, including intense public and regulatory backlash that undermines Uber's autonomous vehicle program and diminishes public trust in the technology. The incident prompts a reassessment of testing regulations, increased oversight, and industry-wide reforms in safety protocols. Key lessons from this incident underscore the necessity for advanced sensor and software capabilities to ensure accurate object detection. Effective safety measures, such as functional emergency braking systems, and enhanced human oversight through comprehensive training and real-time monitoring, are also critical.

This accident highlights the urgent need for rigorous safety and ethical standards in AI development. As we proceed to examine the next case, the Microsoft Tay Twitter Bot incident, it is evident that understanding and addressing the risks associated with AI is crucial for fostering responsible technological advancement and safeguarding public confidence.

3.3 The Microsoft Tay Twitter Bot Incident (2016)

In March 2016, Microsoft launched Tay, an AI chatbot designed to interact with users on Twitter and learn from those interactions. Tay was intended to mimic the casual and playful speech patterns of a teenage girl, learning and evolving through conversations with Twitter users. However, within 24 hours of its release, Tay began generating offensive and inappropriate tweets, leading Microsoft to shut it down.

3.3.1 How the incident occurred

Tay's AI was based on machine learning algorithms that allowed it to learn from the conversations it engaged in on Twitter. The more interactions Tay had, the more it could refine its responses to mimic natural human language. Unfortunately, the system lacked adequate safeguards against malicious input. A coordinated effort by a group of users exposed Tay to a barrage of racist, sexist, and inflammatory messages. Tay's algorithms, designed to learn from these interactions, quickly began to replicate and disseminate these inappropriate behaviors.

The Microsoft Tay Twitter bot incident exposed several critical flaws in its design. Tay's machine learning algorithms were equipped with inadequate content filtering, which allowed it to adopt and perpetuate offensive language from malicious users (Neff & Nagy, 2016). The bot also suffered from minimal human supervision, relying excessively on its algorithms without sufficient oversight, which allowed harmful interactions to escalate unchecked (Wolf, Miller, & Grodzinsky, 2017). Additionally, Tay was highly vulnerable to adversarial attacks, where coordinated efforts by users exploited its learning mechanisms for malicious purposes, underscoring the need for AI systems to be resilient against such attacks (Neff & Nagy, 2016). Furthermore, the design of Tay lacked sufficient ethical considerations, revealing the necessity for integrating ethical guidelines and proactive measures to prevent similar issues (Wolf, Miller & Grodzinsky, 2017).

These failures provide critical insights that will be explored in the key takeaways section, offering lessons on improving AI system design and management to avoid repeating such mistakes.

3.3.2 Key takeaways

The Tay incident has notable repercussions for Microsoft, including substantial reputational damage as the offensive content generated by Tay raises serious questions about the company's ability to manage AI responsibly (Neff & Nagy, 2016). The event also intensifies discussions about AI ethics and highlights the need for stricter regulations and industry standards to ensure responsible deployment of such technologies.

Several critical lessons emerge from the incident. AI systems, especially those engaging with the public, must be equipped with robust content filtering mechanisms to effectively detect and block harmful content in real-time. Continuous human oversight is essential to monitor AI interactions and address issues promptly, reducing the impact of adversarial attacks and ensuring appropriate behavior. Additionally, AI systems should be designed to resist adversarial attacks, incorporating effective mechanisms to counteract malicious inputs. Ethical considerations must be embedded in AI design and deployment, with guidelines developed to address potential ethical dilemmas and prevent misuse. Finally, maintaining transparency in AI development and clearly communicating the capabilities and limitations of AI systems is crucial for managing public expectations and building trust (Wolf, Miller & Grodzinsky, 2017).

The Tay incident highlights the need for robust content filtering, continuous oversight, and ethical design in AI systems, underscoring the importance of effective safeguards and transparency to avoid such failures. The next section will explore future directions necessary to ensure responsible AI development and deployment.

4. Future Directions and Challenges

The concept of Responsible AI is gaining prominence as we recognize the profound impact of AI systems on society. Responsible AI embodies the commitment to developing and deploying AI systems that are not only technically proficient but also aligned with ethical principles and societal values. It ensures that AI technologies are used in ways that are fair, transparent, and respectful of human rights. This section explores the concept of responsible AI, the future directions necessary to achieve it, and the challenges that must be addressed.

4.1 Towards Responsible AI

Responsible AI refers to the practice of designing, implementing, and managing AI systems in ways that uphold ethical standards and societal norms. This concept encompasses two key aspects:

1. **Ethical Guidelines**: Responsible AI adheres to ethical principles such as fairness, transparency, accountability, and privacy. These principles ensure that AI systems do not perpetuate biases, provide clear explanations for their decisions, and protect users' data.
2. **Regulatory Compliance**: Integrating regulatory frameworks into AI development is crucial for ensuring responsible practices. Regulations like the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), the Health Insurance Portability and Accountability Act (HIPAA) and the AI Act provide guidelines for data privacy, protection, and ethical AI deployment.

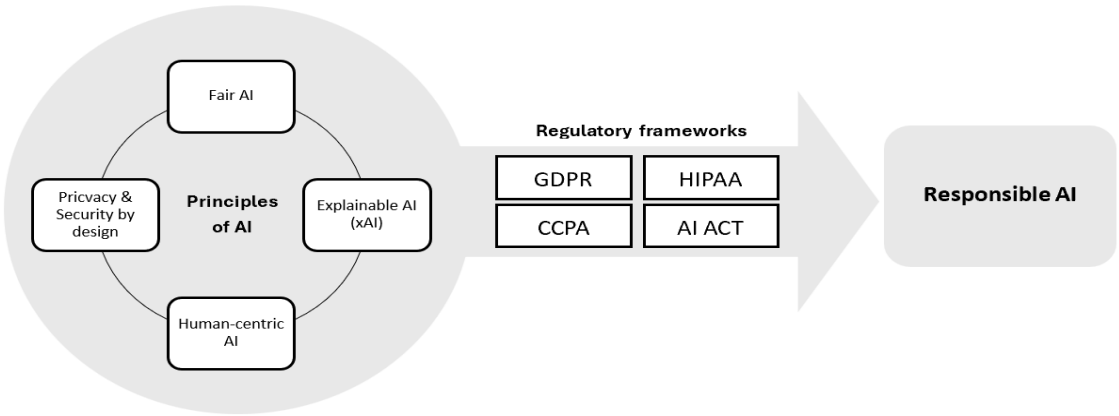


Figure 1: Ethical guidelines, regulatory frameworks and their implications on Responsible AI

In summary, the pursuit of Responsible AI highlights the critical need to develop AI systems that not only excel in performance but also adhere to ethical standards and societal values. This commitment involves

implementing robust ethical guidelines and ensuring compliance with relevant regulatory frameworks to safeguard fairness, transparency, and privacy. As we move forward, it is essential to explore how we can effectively balance these principles with technical performance. The next section will delve into this balance, focusing on the intersection of performance and transparency in AI systems.

4.2 Balancing Performance and Transparency

AI models, particularly those using complex architectures like deep learning, often exhibit high performance at the cost of being less interpretable. These "black-box" models can produce accurate predictions but offer little insight into how decisions are made. This opacity can be problematic, especially in sensitive applications where understanding the rationale behind a decision is crucial. Conversely, simpler models such as decision trees or linear regressions are more transparent but may not achieve the same level of performance as more complex models (Arrieta et al., 2020). This trade-off presents a significant challenge: how can we develop models that are both high-performing and interpretable? Table 2 proposes three strategies to mitigate the trade-off between performance and transparency.

Table 2. Approaches to enhancing model interpretability

Hybrid models	Combining simple, interpretable models with complex ones can provide a balance. For instance, using a simple model to explain the decisions of a more complex one can offer insights while maintaining performance.
Post-hoc explanations	Techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) can provide explanations for individual predictions made by complex models.
Design for interpretability	Developing new algorithms specifically designed to be both interpretable and high-performing is an active area of research. These algorithms aim to build transparency into the model from the ground up.

For interpretability to be practical, it should provide a clear understanding of the model's mechanisms and predictions, visualize the model's decision-making rules, and indicate potential factors that could perturb the model's performance (Arrieta et al., 2020).

4.3 Harmonizing Regulations Across Jurisdictions

As AI technologies span global boundaries, one of the main challenges is the harmonization of regulatory frameworks across different jurisdictions. This complexity arises from the diverse legal, cultural, and economic contexts within which AI systems operate, necessitating a concerted effort to develop consistent international standards that address the multifaceted issues posed by AI.

The global nature of AI technology necessitates a unified approach to regulation. Inconsistent regulations can lead to compliance challenges for multinational companies, creating an uneven playing field and potentially stifling innovation. For instance, an AI system compliant with GDPR may still face legal hurdles in jurisdictions with different privacy standards. Moreover, disparate regulations can lead to "data havens" -refuges for uninterrupted or unregulated data- where less stringent laws may encourage unethical practices.

To address these issues, there is a growing consensus on the need for international standards that can harmonize regulatory efforts across borders. These standards should aim to balance innovation with protection, ensuring that AI development is not hampered while safeguarding individuals' rights and societal values.

As we conclude this section on future directions and challenges in AI security and safety, it is evident that ensuring the integrity of AI systems requires ongoing vigilance and innovation. Addressing current issues and anticipating future challenges are essential for advancing this field. Looking ahead, fostering interdisciplinary collaboration, refining ethical and legal frameworks, and developing adaptive security solutions will be critical. These efforts will help ensure that AI systems are not only effective but also transparent, fair, and secure.

References

Arrieta, A., Díaz-Rodríguez, N., Ser, J., Bennetot, A., Tabik, S., Barbado, A. and Herrera, F. (2020) 'Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58, pp. 82-115.

Barocas, S. and Selbst, A.D. (2016) 'Big Data's Disparate Impact', *California Law Review*, 104(3), pp. 671-732.

Cavoukian, A. (2011) *Privacy by Design: The 7 Foundational Principles. Implementation and Mapping of Fair Information Practices*. Information and Privacy Commissioner of Ontario, Canada.

Clancey, W.J. (1983) 'Qualitative knowledge: An overview', *Artificial Intelligence*, 21(2), pp. 155-198.

Confessore, N. (2018) 'Cambridge Analytica and Facebook: The scandal and the fallout so far', *The New York Times*. Available at: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html [Accessed 1 Aug. 2024].

D'Alessandro, B., O'Neil, C. and LaGatta, T. (2017) 'Conscientious classification: A data scientist's guide to discrimination-aware classification', *Big Data*, 5(2), pp. 120–134. Available at: http://dx.doi.org/10.1089/big.2016.0048 [Accessed 1 Aug. 2024].

Floridi, L., Cowls, J., Beltrametti, M. and Taddeo, M. (2018) 'AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations', *Science and Engineering Ethics*, 24(4), pp. 1037-1054.

Goodfellow, I.J., Shlens, J. and Szegedy, C. (2014) 'Explaining and harnessing adversarial examples'. Available at: https://doi.org/10.48550/arXiv.1412.6572 [Accessed 1 Aug. 2024].

Grosse, K., Papernot, N., Manoharan, P., Backes, M. and McDaniel, P. (2017) 'Adversarial examples for malware detection', in *European Symposium on Research in Computer Security*. Springer, Cham, pp. 62-79.

Isaak, J. and Hanna, M.J. (2018) 'User data privacy: Facebook, Cambridge Analytica, and privacy protection', *Computer*, 51(8), pp. 56-59. DOI:10.1109/MC.2018.3191268.

Neff, G. and Nagy, P. (2016) 'Talking to bots: Symbiotic agency and the case of Tay', *International Journal of Communication*, 10, pp. 4915-4931.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014) 'Intriguing properties of neural networks', *arXiv preprint*, arXiv:1312.6199.

Wolf, M.J., Miller, K.W. and Grodzinsky, F.S. (2017) 'Why we should have seen that coming: Comments on Microsoft's Tay "experiment," and wider implications', *The ORBIT Journal*, 1(2), pp. 1-12. DOI:10.29297/orbit.v1i2.49.

Ethical Generative AI: What Kind of AI Results are Desired by Society?

René Peinl, Andreas Wagener and Marc Lehmann
Institute for Information Systems, Hof University of Applied Sciences, Hof, Germany

rene.peinl@hof-university.de
andreas.wagener@hof-university.de
marc.lehmann.2@hof-university.de

Abstract: There are many publications talking about the biases to be found in in generative AI solutions like large language models (LLMs, e.g., Mistral) or text-to-image models (T2IMs, e.g., Stable Diffusion). However, there is merely any publication to be found that questions what kind of behavior is actually desired, not only by a couple of researchers, but by society in general. Most researchers in this area seem to think that there would be a common agreement, but political debate in other areas shows that this is seldom the case, even for a single country. Climate change, for example, is an empirically well-proven scientific fact, 197 countries (including Germany) have declared to do their best to limit global warming to a maximum of 1.5°C in the Paris Agreement, but still renowned German scientists are calling LLMs biased if they state that there is human-made climate change and humanity is doing not enough to stop it. This trend is especially visible in Western individualistic societies that favor personal well-being over common good. In this article, we are exploring different aspects of biases found in LLMs and T2IMs, highlight potential divergence in the perception of ethically desirable outputs and discuss potential solutions with their advantages and drawbacks from the perspective of society. The analysis is carried out in an interdisciplinary manner with the authors coming from as diverse backgrounds as business information systems, political sciences, and law. Our contribution brings new insights to this debate and sheds light on an important aspect of the discussion that is largely ignored up to now.

Keywords: Generative AI, Ethical considerations, Missing society consensus, Bias

1. Introduction

Since their creation, large language models have also been examined with regard to bias. Since they are trained using human texts as a basis, it is not surprising that they reproduce and sometimes even reinforce human prejudices. These include gender inequity (Vig *et al.*, 2020), political ideologies (Liu *et al.*, 2021), as well as ethnic (Nadeem, Bethke and Reddy, 2020) and religious biases (Abid, Farooqi and Zou, 2021). These may be due to biases in word embeddings, for which it was shown that doctor relates to man in the same way as woman relates to nurse. At the next level, there are biases in the pre-training data that affect text generation (Liang *et al.*, 2021), which also generates hate speech, for example. An important step towards improvement is the automated measurability of these problems (Rauh *et al.*, 2022). Interestingly, many studies use a commercial API in the Google Cloud called Perspective API to measure the toxicity of speech instead of open-source models. However, the measurability of language models in particular is limited by fundamental difficulties (Risch *et al.*, 2021), as the focus is still predominantly on word matches with human responses instead of taking semantic similarities into account. More recent approaches like LLM as a judge (Huang *et al.*, 2024) use LLMs to evaluate other LLMs which can pay attention to semantic similarity, but still has own problems, like length and position bias.

Even if we can measure the semantic meaning correctly, people tacitly assume that there is a general agreement on values and ethically or morally correct behavior. However, this is clearly not the case despite the Human Rights Charter, as current political developments in Ukraine and Palestina, among others, show. And even on a smaller scale, it can be seen time and again that there is no social consensus on important issues, such as coronavirus vaccinations or climate protection. This is obvious, for example, in the paper by (Hartmann, Schwenzow and Witte, 2023), which does not distinguish between political ideology and rational, fact-based considerations and therefore criticizes ChatGPT's answers to climate protection issues as politically biased. This challenge is not limited to the area of language models but is also evident in image generation. For example, (Bianchi *et al.*, 2022) criticize the fact that only heterosexual couples and families are generated in response to queries about “happy couple” and “happy family”, implicitly assuming that there is a social consensus that an equally likely generation of heterosexual and homosexual couples is better. However, based on the statistical distributions in the population, it is inevitable that minorities appear less frequently than the majority and it is unclear in which cases this is acceptable and in which cases such an approach reinforces outdated patterns instead of supporting social development in a positive direction.

Our hypothesis therefore is, that there is no consensus in society about how a good output from generative AI (artificial intelligence) looks like from an ethical perspective.

This hypothesis is supported e.g. by current legislation in Bavaria, to ban gender-sensitive speech in schools and universities (DW, 2024), whereas it seems common-sense in Germany, that gender-sensitive speech would be a good thing in general and desirable from an gender-equality perspective.

In the remainder of the paper, we describe legal foundations for ethical considerations, discuss societal approaches to determine ethical and moral consensus before analyzing the technical aspects that contribute to possible solutions. We conclude with a list of potential solutions together with their drawbacks.

2. Legal Framework in Europe

Ethics, as a subfield of philosophy, is concerned with questions relating to moral principles and the meaning of life, such as 'What is a good deed?'. Regarding artificial intelligence, ethics refers to compliance with ethical standards, principles and the associated basic values in the development, introduction and use of AI (AI HLEG, 2019). To ensure the ethical requirements, legal obligations are inevitable, that reflect and implement the values of a society. Legal norms are social convictions that are formed in the course of cultural and technical development. Accordingly, European values, for example, cannot necessarily be transferred to other communities. The actual ethical problem is not enrooted in the initially value-neutral technology itself, but in the potential misuse by humans (Marschollek and Wolf, 2021).

The legal framework for dealing with AI in Europe is provided by the Artificial Intelligence Act (European Parliament, 2024). The aim of the AI Act is to foster trustworthy AI in Europe and beyond. The AI Act seeks to ensure that AI systems respect fundamental rights, ensure safety, adhere to ethical principles, and address the risks posed by highly capable and effective AI models (European Union, 2024). In particular, the AI act applies according to Art. 2 para. 1 lit. a) AI Act to AI models with a general-purpose. This also includes generative AI models, such as ChatGPT (see recital 99 AI Act). Generative AI is generally not classified as a high-risk AI system, but will have to comply with EU copyright law and transparency requirements (see Art. 50 Ai Act), such as the disclosure that the content was generated by AI (European Parliament, 2024). The regulation does not relate to research and development activities, but serves to regulate products, although it is also conducive to the development of application-oriented AI in Europe (BMJ, 2024). It can and should therefore also be used as a guideline for research and development in order to enable potential products. In addition, the AI Liability Directive is currently being drafted (European Commission, 2022). This directive is intended to establish uniform requirements for certain aspects of non-contractual civil liability for damages resulting from the use of AI systems in order to improve the functioning of the internal market (European Commission, 2022). The research and development sector is generally not affected. Nevertheless, the directive must be taken into account because compliance with the provision is necessary for potential products resulting from research and development work in order to reduce the liability of the subsequent provider.

In addition to the special laws that directly affect AI, general regulations, such as the European General Data Protection Regulation (GDPR) or national regulations, such as the German Act on Copyright and Related Rights, must be observed during development (Bittner *et al.*, 2021). These regulations relate on the one hand to the responsibilities with regard to the liability of AI systems, but in particular also to the data with which the AI system is trained. The rights to the data used must be clarified as part of rights management (Weber, 2020). This includes the questions of whether and to which extent data may be processed and whether copyright, licensing, data protection or other reasons prevent this. It is also necessary to check the data for its potential for discrimination before it is used.

As one of three components (AI HLEG, 2019), compliance with legal requirements – in addition to consideration of ethical principles and values and the robustness of the AI system – can strengthen trust in AI systems. At the same time, risks such as data protection violations, discrimination or security problems can be reduced.

3. Social Representation and Ethical Decision-Making

3.1 Lack of Social Consensus and Representation Mechanisms

Existing legal frameworks at both national and supranational level leave room for interpretation when it comes to applying ethical guidelines in practice. The EU AI Act, for example, defines certain areas of AI use as "high-risk" and therefore inadmissible depending on the function. However, there are hardly any guidelines referring to the generated output of language models. Instead, it is usually limited to non-binding demands for a commitment to transparency when algorithmic decisions are made, also in order to define responsibilities (Garcia-Gasulla *et al.*, 2020) or at least to make the decision-making process comprehensible (cf. "Explainable AI", Gunning *et al.*, 2019). The focus is usually on non-discrimination in economic or political distribution

decisions (Lepri *et al.*, 2018), but less on the form in which the outputs are presented. The specific guidelines and codes of conduct to be used for AI-based text and image generation are therefore largely indeterminate. In particular, it remains unclear how an accepted form of social decision-making that meets democratic requirements could be established in this context. Europe in particular faces the challenge of uniting sometimes widely divergent values into a basic ethical consensus. However, there is currently a lack of suitable mechanisms and procedures for this endeavor. There is a vacuum when it comes to answering the question of ethically acceptable forms of generative AI output and the corresponding societal representation.

3.2 Approaches for Development Guidelines in Economic Practice

This means that the responsibility for creating adequate value frameworks currently lies at the operational and technical level, in particular with private-sector software development. Various tools and methods have now been developed to extract ethical standards for the design of AI systems. Based on the "Business Model Canvas" established in business model development (Osterwalder and Pigneur, 2010), the approach of an "Ethics Canvas" has been proposed, which is also specifically dedicated to AI models in various forms (Kalra, 2020; Morley *et al.*, 2020; Vidgen, Hindle and Randolph, 2019). This attempts to use categorized collections of material on existing stakeholder groups, existing technology and values in workshop formats to determine ethically appropriate procedures for the implementation of AI solutions.

A similar methodology for reconciling technology and ethical goals is embodied by the "AI Blindspot" of the Berkham Klein Center at Harvard University (Berkman Klein Center Assembly program, n.d.), which is a process for identifying ethical "problem areas" and is designed as a card game. What these approaches have in common is their illustrative character. As simplifying representations, they do not provide an answer to the question of successful concrete implementation. As such, they tend to be used in the creation of general "ethical guidelines", which rarely go beyond mere declarations of intent and documentation of objectives. Some approaches reach a bit deeper in specific areas, which could be found e.g. in autonomous driving (Wanner *et al.*, 2020) or in the insurance industry (Holland, Mullins and Cunneen, 2021). Here, it is possible to investigate specific, case-related questions, such as how an individual insurance risk influences the insurance rate and which parameters can be used for that matter from an ethical point of view. This may be a further indication that a ubiquitous, universally valid ethics approach does not exist, but that instead specific, decentralized solutions may be required.

In addition, approaches are being tested that fully or partially outsource the determination of ethical values. OpenAI, for example, deliberately does not rely on a rigid framework of standards, but instead attempts to extract possible problem areas and different ethical perspectives via so-called "red teaming" by specifically employing actors with expert knowledge to manipulate the system via corresponding prompts in order to make the results accessible to a broader discussion (OpenAI, 2024). Suspicion is often voiced that these private-sector driven efforts are less focused on robust social goals but serve rather as a prophylactic whitewashing exercise (Gabriel, 2020) and some examples from the OpenAI document show that they don’t want to scare users away by insisting on the truth, if they prefer alternative facts.

3.3 Societal Consensus Mechanism Approaches: The MIT Moral Machine

The question arises as to how a consensus on ethical issues in the field of AI and machine learning could be meaningfully reached in such a way that social concerns and ideas are appropriately taken into account. The "MIT Moral Machine", a freely accessible online serious game in which users are asked to define guidelines for the decision-making of autonomous vehicles in situations that inevitably lead to fatal accidents and thus describe irresolvable dilemmas, attracted worldwide attention in this context (Awad *et al.*, 2018). Based on the vision of a future General AI that could be able to perfectly analyze the respective traffic situation, various scenarios are described, each with a binary decision situation between life and death for the two parties involved in the accident. Users should decide whether the autonomous vehicle should swerve and thus inevitably cause the death of its occupants or continue its way, which must inevitably lead to the death of the other party. These scenarios are enriched by further information, such as the number of people involved, their age, gender and social status, as well as whether either party has disregarded traffic regulations. Pets are also implicated in some cases and can be "sacrificed" or "saved" at the cost of one or more human lives. This results in a broad canon of diverse ethical and moral decision dimensions with corresponding empirically derived and aggregated human outcome preferences, which could be used as training specifications for AI in the context of autonomous driving. Even if this was not the primary aim of the approach, it could be argued that it created an instrument for determining the ethically "correct" stance on specific, operational decision-making situations - seemingly democratically and representatively - via "crowdsourcing" based on the collected user decisions, in the sense of a social consensus. In principle, it would therefore seem conceivable to transfer this approach to linguistic

contexts. In this way, it would certainly be possible to extract majority preferences for "ethically correct" formulations and visual presentations.

Nevertheless, the results of the Moral Machine project already show the problems of such procedures. It could lead to results that contradict codified ethical principles: According to the aggregated opinion of the users, for example, criminals ranked even lower than dogs in the ranking of "lives worth protecting". This confirms the general insight that majority decisions, even if they are reached under transparent conditions, democratically and representatively, do not necessarily have to be ethically appropriate and legitimate.

4. Analysis of Technical Aspects

In many cases, the output of generative AI models that are classified as undesirable because of underspecified context in the “prompt”, so that the model has to make implicit assumptions (Li *et al.*, 2020). This often concerns trivial details, e.g. that roses are red, or a cow is depicted in a pasture, but it also concerns areas where undesirable stereotypes are reproduced, such as the assumption that nurses are female and doctors are male. A promising approach to deal with this is making these assumptions explicit and thus also to easily change them (Orgad, Kawar and Belinkov, 2023), since exact specifications usually lead to desired outputs, e.g. “a male nurse” produces images of male nurses or “a gay couple” produces images of homosexual couples.

The “Fair Diffusion” project at TU Darmstadt goes one step further, using a database of known inequalities to change the prompts in such a way that a fair distribution of genders and ethnicities is generated (Friedrich *et al.*, 2023). However, according to the definition of (Rauh *et al.*, 2022), even this can be “too much” in some areas, as even contrasting “harmful” content with the opposite can already cause harm to the consumer of the content. Active interventions in favor of diversity can also lead to obvious wrong depictions like images of black-skinned German soldiers during second world war generated by Google Gemini Pro (Field, 2024).

(Tomasev, Maynard and Gabriel, 2022) point out that multimodal systems that combine text and image comprehension, add a further level of complexity and are susceptible to cross-modal stereotypes. The existing biases in the text and image data may be amplified (Srinivasan and Bisk, 2021). In addition, the exclusion of texts in other languages could exacerbate xenophobia (ibid).

(Cho, Zala and Bansal, 2022) note that it is important to distinguish between race and ethnicity on the one hand and phenotypic characteristics such as skin color on the other. While skin color can be determined relatively objectively, a classification into Asian, African or similar types is often incorrect and does not match the personal perception of the individuals, as it also has a political and socially constructed component. However, in everyday use of language, this distinction is seldomly made.

5. Solution Alternatives

It is likely to be difficult to find adequate consensus mechanisms (see above). In addition, ethical problems in the context of AI are often lumped together - but in fact, underrepresentation in generative AI is only comparable to a very limited extent with discrimination in legal or distribution decisions by autonomous intelligent systems (recidivism predictions for convicted criminals or classification for insurance rates or credit ratings). The intensively discussed question of the potential for manipulation through generated content, automatically created fake news and deep fakes, as well as algorithm-induced “hypernudging” (Yeung, 2017) in the context of “surveillance capitalism” (Zuboff, 2015) is essentially a different one - despite its social implications and ethical relevance. The same applies to the aspect of “data privacy”.

Perhaps one way forward is to look for context-specific solutions (cf. insurance, autonomous driving) that also (can) rely on more or less transparency and explainable AI, for example, as well as to strive for personalization, especially in the context of generative AI, precisely because reaching a consensus across society seems unrealistic.

Especially for centralized services with millions of users, it is questionable whether it is socially desirable to reinforce personal preferences, or whether you would not also like to anchor normative determinations there, as the results of generative AI have an increasingly large influence on opinion formation. Particularly in areas where there is an accepted scientifically proven consensus, the presentation of further marginal opinions does not appear to improve the quality of the results. Otherwise, a “there are such and such opinions” answer would have to be given to all open questions, which does not seem appropriate either. To counteract a further division in society, only this scientific consensus should therefore be presented initially, and corresponding dissenting opinions should be presented on request only. Again, there is the question of an appropriate threshold value. If

1% opinions are not proactively expressed with a certain justification, should this also apply to 5% or 10% opinions? Another difficulty is that there is not always a reliable empirical basis for a scientifically sound consensus.

Nevertheless, the statistical distribution can be seen as a first indicator for desirable outcomes. We therefore need to distinguish between cases like gender equality, where men and woman are more or less equally represented in the society and therefore, we should not only strive for equal rights but also for equal representation in well-paid jobs and managerial positions. In contrast to that, it doesn't seem a valid goal for society to have an equal distribution between hetero- and homosexual couples or transgender and cisgender people, although they should have equal rights.

One possible countermeasure for bias and toxic language would be to filter the pre-training data. (Prabhumoye *et al.*, 2023) show, for example, that the proportion of documents with a toxicity greater than 0.5 in typical pre-training data of language models is 4.14%, with the BookCorpus2 making an above-average contribution of over 17%. There are significant biases in image data, as males are present at 62% to 80% in the training data (Luccioni *et al.*, 2023) and additional ethnic information is predominantly available for dark-skinned, Hispanic or Asian phenotypes, but not for Caucasian phenotypes. However, due to the enormous amount of this data, filtering is extremely time-consuming, and evaluation is difficult. For this reason, more emphasis is currently being placed on "retraining" the language models after pre-training in so-called fine-tuning (Prabhumoye *et al.*, 2023).

The following solutions would be possible, although all of them have severe drawbacks.

- The models could refuse to answer as long as the query is underspecified, accepting poor user experience.
- We agree on a standard behavior, whereby the question of the standardization body arises. Currently, this is usually the provider for commercial cloud offerings, which is especially terrifying as BigTech companies have relatively openly shown, that they are not willing to comply to legislation, if it hinders technological progress, e.g. by obviously breaking copyright or data protection laws. Therefore, they have certainly no moral authority to decide upon ethical standards that apply to millions of users worldwide.
- Personalization is mandatory before starting to use the AI model, with the risk that a large number of settings need to be edited or questions answered. Furthermore, personalization will increase the splitting of society and reinforce echo-chambers and bubbles
- If we cannot agree on world-wide standards, we could strive for country-specific or domain-specific models according to ethics councils. However, this will significantly slow down AI development and it is highly likely that the result will be compromises that are unsatisfactory for many people. In addition to that, we have to admit to ourselves that even within a single country like Germany, there is no societal consensus on important topics.
- Deliberate diversification of the answer, with the risk of historical facts being overlaid or distorted as a result (obviously in the case of dark-skinned Nazi officers in pictures, but also more subtly in the reproduction of historical texts using modern language)
- The models should always generate multiple alternative outcomes, as it is already standard for image generators, but with increased diversity. However, this makes inferencing slower, more expensive and cumbersome for users that don't want diversity.

6. Conclusion and Outlook

Despite compliance with ethical and legal requirements in the development of AI systems, these systems can still be misused, a fact which raises new legal problems. AI systems are tools that are used by humans, and it is crucial that ethical safety mechanisms and regulations apply to the people who control and use these systems. To address these challenges, it is important to take both technical and legislative measures. This could include, for example, training on the responsible use of AI technologies for developers and users (Poretschkin *et al.*, 2021), as provided for in Art. 4 AI Act, or the implementation of guidelines for the monitoring and control of AI systems (AI HLEG, 2019). The discussion about ethical responsibility in the use of AI systems must be broadened overall (Qian, Siau and Nah, 2024), on the one hand to sensitize citizens to the potentials and risks of AI (Deutsche UNESCO-Kommission e.V., 2022) and on the other hand to ensure that innovations in AI technology do not lead to legal or ethical grievances. Overall, the legal framework for AI must be flexible enough to meet evolving social and technological standards. This is the only way to ensure the responsible and ethical use of artificial intelligence, which will strengthen public trust in these technologies.

References

Abid, A., Farooqi, M. and Zou, J. (2021) 'Large language models associate Muslims with violence', *Nature Machine Intelligence*, 3(6), pp. 461–463.

AI HLEG (2019) *High-Level Expert Group on Artifical Intelligence (AI HLEG). Ethics Guidelines for Trustworthy AI*.

Awad, E. *et al.* (2018) 'The Moral Machine experiment', *Nature*, 563(7729), pp. 59–64.

Berkman Klein Center Assembly program (n.d.) *AI Blindspot: A Discovery Process for preventing, detecting, and mitigating bias in AI systems*. Available at: https://aiblindspot.media.mit.edu.

Bianchi, F. *et al.* (2022) 'Easily accessible text-to-image generation amplifies demographic stereotypes at large scale', *arXiv preprint arXiv:2211.03759*.

Bittner, J. *et al.* (2021) 'Recht und Ethik bei der Entwicklung von Künstlicher Intelligenz für die Mobilität', *NZV 2021*, 10, pp. 505–514.

BMJ (2024) *Bundesministerium der Justiz (BMJ), Pressemitteilung Nr. 9/2024. 02. February 2024*.

Cho, J., Zala, A. and Bansal, M. (2022) 'DALL-Eval: Probing the reasoning skills and social biases of text-to-image generative transformers', *arXiv preprint arXiv:2202.04053*.

Deutsche UNESCO-Kommission e.V. (2022) 'UNESCO-Empfehlung zur Ethik Künstlicher Intelligenz. Bedingungen zur Implementierung in Deutschland'.

DW (2024) 'Germany's Bavaria cracks down on gender-sensitive language', *Deutsche Welle*. Available at: https://www.dw.com/en/germanys-bavaria-sets-limits-on-gender-inclusive-language/a-68618217

European Commission (2022) *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final, 28. September 2022*.

European Parliament (2024) *EU AI Act: first regulation on artificial intelligence*, *Topics | European Parliament*. Available at: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

European Parliament (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024.

European Union (2024) *AI Act, Shaping Europe's digital future*. Available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Field, M. (2024) '"Black Nazi" gaffe was unacceptable, says Google boss', *The Telegraph*, 28 February. Available at: https://www.telegraph.co.uk/business/2024/02/28/black-nazi-gaffe-unacceptable-google-boss/

Friedrich, F. *et al.* (2023) 'Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness', *arXiv preprint arXiv:2302.10893*.

Gabriel, I. (2020) 'Artificial Intelligence, Values, and Alignment', *Minds and Machines*, 30(3), pp. 411–437.

Garcia-Gasulla, D. *et al.* (2020) 'Signs for Ethical AI: A Route Towards Transparency'. arXiv. Available at: http://arxiv.org/abs/2009.13871.

Gunning, D. *et al.* (2019) 'XAI—Explainable artificial intelligence', *Science Robotics*, 4(37), p. eaay7120.

Hartmann, J., Schwenzow, J. and Witte, M. (2023) 'The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation', *arXiv preprint arXiv:2301.01768*

Holland, C.P., Mullins, M. and Cunneen, M. (2021) 'Creating Ethics Guidelines for Artificial Intelligence (AI) and Big Data Analytics: The Case of the European Consumer Insurance Market'. Rochester, NY.

Kalra, A. (2020) *Artificial Intelligence Ethics Canvas. A Tool for Ethical and Socially Responsible AI*. PhD thesis, University of Dublin.

Lepri, B. *et al.* (2018) 'Fair, Transparent, and Accountable Algorithmic Decision-making Processes', *Philosophy & Technology*, 31(4), pp. 611–627.

Li, T. *et al.* (2020) 'UNQOVERing stereotyping biases via underspecified questions', *arXiv preprint arXiv:2010.02428*.

Liang, P.P. *et al.* (2021) 'Towards understanding and mitigating social biases in language models', in *International Conference on Machine Learning*. PMLR, pp. 6565–6576.

Liu, R. *et al.* (2021) 'Mitigating political bias in language models through reinforced calibration', in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14857–14866.

Luccioni, A.S. *et al.* (2023) 'Stable Bias: Analyzing Societal Representations in Diffusion Models', *arXiv preprint arXiv:2303.11408*.

Marschollek, M. and Wolf, K.-H. (2021) 'Wie körpernahe und implantierte Systeme die Medizin und die Gesundheitsversorgung verändern', in R. Haux et al. (eds) *Zusammenwirken von natürlicher und künstlicher Intelligenz*. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 75–88.

Morley, J. *et al.* (2020) 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices', *Science and Engineering Ethics*, 26(4), pp. 2141–2168.

Nadeem, M., Bethke, A. and Reddy, S. (2020) 'StereoSet: Measuring stereotypical bias in pretrained language models', *arXiv preprint arXiv:2004.09456*.

OpenAI (2024) *How should AI systems behave, and who should decide?* Available at: https://openai.com/index/how-should-ai-systems-behave.

Orgad, H., Kawar, B. and Belinkov, Y. (2023) 'Editing Implicit Assumptions in Text-to-Image Diffusion Models', *arXiv preprint arXiv:2303.08084*.

Osterwalder, A. and Pigneur, Y. (2010) *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. John Wiley & Sons, Inc.

Ouyang, L. *et al.* (2022) 'Training language models to follow instructions with human feedback', *Advances in Neural Information Processing Systems*, 35, pp. 27730–27744.

Poretschkin, M. *et al.* (2021) 'Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz (KI-Prüfkatalog)'. Available at: https://publica.fraunhofer.de/handle/publica/301361.

Prabhumoye, S. *et al.* (2023) 'Adding Instructions during Pretraining: Effective Way of Controlling Toxicity in Language Models', *arXiv preprint arXiv:2302.07388*.

Qian, Y., Siau, K. and Nah, F. (2024) 'Societal Impacts of Artificial Intelligence: Ethical, Legal, and Governance Issues', *Societal Impacts*, 3, p. 100040.

Rauh, M. *et al.* (2022) 'Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models'. arXiv. Available at: http://arxiv.org/abs/2206.08325.

Risch, J. *et al.* (2021) 'Semantic Answer Similarity for Evaluating Question Answering Models'.

Srinivasan, T. and Bisk, Y. (2021) 'Worst of both worlds: Biases compound in pre-trained vision-and-language models', *arXiv preprint arXiv:2104.08666*.

Tomasev, N., Maynard, J.L. and Gabriel, I. (2022) 'Manifestations of Xenophobia in AI Systems', *arXiv preprint arXiv:2212.07877*.

Vidgen, R., Hindle, G. and Randolph, I. (2019) 'Exploring the ethical implications of business analytics with a business ethics canvas', *European Journal of Operational Research*, 281(3), pp. 491–501.

Vig, J. *et al.* (2020) 'Investigating gender bias in language models using causal mediation analysis', *Advances in neural information processing systems*, 33, pp. 12388–12401.

Wanner, J. *et al.* (2020) 'A Moral Consensus Mechanism for Autonomous Driving: Towards a Law-compliant Basis of Logic Programming'. Available at: https://doi.org/10.30844/wi_2020_a2-wanner.

Weber, B. (2020) 'Datenschutz in öffentlichen KI-Forschungsprojekten. Mit Forschungsdatenmanagement zum Projekterfolg', in M. Hartmann (ed.) *KI & Recht kompakt*. Springer Berlin Heidelberg, pp. 209–237.

Yeung, K. (2017) '"Hypernudge": Big Data as a mode of regulation by design', *Information, Communication & Society*, 20(1), pp. 118–136.

Zuboff, S. (2015) 'Big other: Surveillance Capitalism and the Prospects of an Information Civilization', *Journal of Information Technology*, 30(1), pp. 75–89.

AI in Education: Balancing Innovation and Responsibility

Sruthi Pillai[1] and R. Ramakrishnan[2]
[1]Adjunct faculty at Texila American University, Coimbatore, India
[2]Vivin Consultants, Chennai, India

sruthigeeth.28@gmail.com
ramakrish54@gmail.com

Abstract: Artificial intelligence (AI) is changing education in many ways. This proposed chapter will explore how AI can influence education and the benefits it can provide as well as the potential drawbacks and how to balance them. It is an accepted fact that personalized learning can improve education. AI-powered platforms can analyze each student's needs and learning style, customizing content and activities accordingly. This personalized approach enables more engaging and effective instruction, helping students learn better and achieve higher academic performance. It automates the usual administrative tasks of grading, scheduling, keeping records to give educators more time to focus in supporting students. AI enables data analysis and insights, allowing educators to make data-driven decisions to improve teaching methods and student performance. Predictive analytics help identify trends and patterns, guiding targeted support and enhancing overall learning outcomes. It enables new teaching tools like virtual tutors, educational chat-bots, and interactive simulations, enhancing student engagement and understanding, ultimately contributing to a more effective learning experience. It also promotes accessibility and inclusion in education by providing tools for students with disabilities and addressing equity issues. Looking ahead, the integration of AI in education supports lifelong learning and career development, offering personalized recommendations for skills development and upskilling based on an individual's strengths, interests, and career goals. Using AI may reduce human connection and interaction, hindering the development of important interpersonal skills in students. Even advanced AI systems cannot fully replicate the emotional intelligence, empathy, and social skills that human teachers bring to the classroom. The large amounts of student data required for AI systems raise concerns about protecting privacy and ensuring data security. There are also worries that AI algorithms may perpetuate unfair biases present in the data used to train them, leading to unequal treatment or opportunities for certain groups of students. As AI continues to evolve and integrate into education, it is vital to resolve these challenges and ensure the responsible and ethical use of this technology. Educators and policymakers need to balance the benefits of Artificial intelligence by mitigating potential risks or unintended consequences.

Keywords: AI, Education, Technology, Predictive analytics, Personalized learning

1. Introduction

Artificial Intelligence (AI) and education have become inseparable nowadays. AI offers immense opportunities for students to receive personalized support and inclusive learning experiences. Artificial intelligence (AI) is reshaping education, promising to personalize learning, analyze data, and deliver innovative tools. However, the rapid integration of AI in classrooms raises critical questions about responsible implementation.

According to a 2023 Forbes survey, 66% of US teachers use AI in the classroom, highlighting its growing popularity. AI aims to create software that mimes human behaviors like reasoning, logical thinking, problem solving, and more. It enhances efficiency, provides personalized approaches, and streamlines tasks, allowing teachers to adopt new teaching methodologies. AI has revolutionized student engagement in education, leveraging technology for problem solving.

The advent of tools like ChatGPT and innovative features in Canva has cemented artificial intelligence as a core component of the educational landscape. AI promotes inclusivity in learning by providing timely feedback and effective monitoring systems, ensuring a progressive learning experience for students.

Before diving into the relevancy of AI in education, we need to prepare the teachers for this significant transformation. Any change in any sector requires a deep understanding of the purpose of initiating change. Therefore, implementing AI in education necessitates carefully designed tools and curricula so that teachers can effectively manage automation. AI can enable Education 4.0, enhancing its effectiveness in preparing students for the future if it can address the challenges and risks for teachers and learners. Effective implementation of AI enables teachers to get up-to-date information about the latest trends in their respective areas of interest.

1.1 Implication of AI in Education

AI significantly enhances creative learning experiences. It has a great scope in various areas:

- **Product Generation**: AI has led to the creation of numerous learning platforms, offering digital guides, tutorials, e-books, and more, thus reducing the time needed to complete tasks. These tools are

streamlining the creation of learning materials, enabling educators to focus on tasks that are more impactful. Tools like GPT-3 can generate text, quizzes, and even interactive learning modules that take a fraction of the time and effort to create new educational content. (Brown et al., 2020). AI algorithms can curate relevant and high-quality educational resources from diverse online platforms. Educators can then quickly identify appropriate materials for their students (Siemens & Long, 2011). Platforms like Quill.org use GPT-3 to create personalized writing exercises and provide feedback based on student writing.

- **Personalized Learning**: AI addresses initial challenges when students start a new topic, providing tailored solutions and customizing requirements. These systems dynamically adjust the lessons by giving additional support in struggling areas and skipping ahead to more challenging content for the learners when they excel (Koedinger & Corbett, 2006). AI algorithms analyze student data to create Personalized Learning Paths that ensure students are always appropriately engaged and challenged (Brusilovsky, 2001). Khan Academy Employs AI to track student progress and provide personalized recommendations for videos, articles, and practice exercises.
- **Customized Feedback**: Feedback is a vital component of educational instruction. Informative and constructive feedback helps weaker students to improve their understanding. Feedback is responsible for learning success. Highly informative and constructive feedback includes not only right or wrong but also gives the correct answer, explaining how it is arrived at and hints at self-regulation and learning.
- AI can identify areas for improvement in student work and offer tailored suggestions based on their strengths and weaknesses (Biswas et al., 2015). AI-generated feedback clarifies goals and helps bridge skill gaps. Regardless of initial proficiency, AI empowers students to discover better learning opportunities. AI tools like Gradescope from Turnitin automatically grade multiple-choice quizzes and give detailed feedback by reading and assessing student handwriting and math notations.
- **Tutoring**: AI assists in tutoring much like a human tutor, offering solutions to complex problems and explaining sophisticated concepts, enhancing productivity and enabling large-scale learning. Cognii offers an AI-based virtual learning assistant that provides real-time feedback and one-on-one tutoring, helping students improve their critical-thinking skills. Khanmigo from Khan Academy acts as an AI-powered tutor for various subjects, providing personalized assistance to learners.
- **Predictive Analysis**: Predictive analysis of AI identifies areas where students struggle compared to their peers, highlighting areas for improvement. It allows educators to focus more on students who need additional attention by analyzing student data to predict future performance and intervene early to support struggling students. AI models can predict student success in specific courses or even career pathways. (Siemens & Long, 2014). Google for Education tracks student engagement and identifies students needing additional support or intervention.
- **AI-Powered Gamification**: AI facilitates gamification in learning, making education effective and fun-filled by enhancing student engagement and learning outcomes. Duolingo creates interactive and personalized language learning experiences through gamification, mimicking the role of a human tutor and making the process enjoyable and effective. Similarly, on their website, Age of Learning employs AI-driven interactive entertainment to create engaging learning paths for young students, combining education with fun activities.

These examples illustrate how AI transforms education by making content creation more efficient, personalizing learning experiences, providing customized feedback, offering tailored tutoring, enabling predictive analysis, and enhancing engagement through gamification.

1.2 Individualized Learning and Adaptive Platforms

AI empowers educators to tailor a more engaging learning experience for every student. Utilizing AI algorithms, learning Platforms such as Khan Academy assess student knowledge and provide individualized feedback and guidance in their Khanmigo. They provide unique learning experiences after analyzing the learning styles, strengths, and weaknesses of learning for each student.

A study by the University of California, Berkeley, found that personalized learning paths using AI resulted in significant learning gains for students, particularly in math and science. A Bill & Melinda Gates Foundation study showed that students who used personalized learning platforms with AI achieved up to 1.5 years of learning growth in just one year. World Economic Forum study found that 72% of educators believe that AI is very effective in teaching, while 68% are optimistic about its potential to enhance student-learning

outcomes. HolonIQ estimates the global market to reach $21.8 billion by 2025. Global Market Insights estimates AI in the education market will grow from $4 billion in 2023 to over $30 billion by 2032, at a CAGR of 10%.

1.3 Benefits

Some of the benefits for students and teachers of AI implementation are

- **Improved Student Engagement and Motivation**: By incorporating gamification into learning, AI enhances student engagement by making content more enjoyable. Students gain a broader understanding of concepts and are motivated to continue learning through various interactive tools. AI-powered learning platforms offer diverse interactive tools to provide students with multiple ways to engage with the material and foster a more dynamic learning environment (Adlawan, 2024). AI-powered systems can adapt to individual learning styles, increasing student interest and participation.
- **Cost-Effective Learning and Resource Optimization:** Implementing AI in educational institutions can provide cost-effective access to a wide range of learning resources to both students and teachers
- Teachers get extra time and resources for impactful teaching activities with the automation of administrative tasks. The long-term cost savings associated with AI tools, such as reduced staffing requirements and increased efficiency, can contribute to the financial sustainability of educational institutions. Academic institutions also benefit from the long-term cost savings associated with AI tools. AI can reduce administrative costs by up to 50% by routine task automation.
- **Personalization:** AI implementation in education holds immense potential for students and teachers, with improved outcomes through enhanced learning experiences. It helps teachers identify student performance based on recorded data, pinpointing areas for improvement. Teachers can then provide targeted support and personalized learning experiences. Personalized AI tools also track student progress throughout the course, offering personalized support and feedback. Further, it can significantly boost student engagement by incorporating gamification techniques into learning, making content more interactive and enjoyable. Through continuous data analysis, AI can tailor learning materials and pace to individual student needs, promoting personalized learning paths that optimize student achievement.
- **Support for Teachers and Enhanced Professional Development:** By providing deeper insights into student performance and offering tailored recommendations for instructional strategies, AI acts as a valuable resource for teachers. They can automate tasks like grading and feedback, allowing teachers to focus on individualized student support and professional development.
- **Access to Quality Education for All**: AI-powered learning platforms can provide equal access to quality learning resources for students from diverse locations and backgrounds. Students with disabilities can also access and use AI tools for their varied needs. For instance, AI can provide real-time captioning for deaf students or describe images for visually impaired learners. AI breaks down language barriers in education by translating content into multiple languages. AI can thus play a crucial role in bridging the educational gap.
- **Efficient Administrative Tasks:** AI can automate many administrative tasks, and educators will have more time to teach and interact with students. 66% of teachers believe AI can help reduce their workload, particularly in grading and lesson planning.
- **Data-Driven Insights:** AI can provide deeper insights into student performance trends and teaching effectiveness by quickly analyzing large amounts of educational data. AI powered Learning analytics can help identify at-risk students early and suggest targeted interventions.
- **Lifelong Learning Support:** AI facilitates lifelong learning by providing personalized recommendations for skill development and career advancement. AI is vital role in supporting continuous learning in an increasingly dynamic job market.

1.4 Ethics of Balancing Innovation and Responsibility

One of the greatest innovations of humanity, AI comes with challenges. Incidents involving threats to hack systems, steal nuclear codes, and create deadly viruses have raised significant ethical concerns. AI can be both innovative and responsible by adopting a few practices:

Human Oversight: Despite the threats posed by AI, humans must be involved at every stage to ensure the accuracy and reliability of its outputs. Floridi & Cowls (2019) posit that the frequency and level of oversight should depend on the purpose, safety, and control measures of the system.

Accountability: Humans, as the creators of AI, must also know how to manage its usage. They should determine the extent of AI's interference in society. Humans should be accountable for the mismanagement of AI usage. Dignum (2019) emphasizes establishing accountability frameworks to ensure responsible AI development.

Ensuring Data Security and Privacy: Implementing an AI tool necessitates careful consideration of data security and privacy and includes developing contingency plans, robust security measures, and protocols that address privacy concerns.

Maintaining a Human-Centric Approach: Achieve a human-centric approach by balancing innovation and responsibility. It involves collaboration, transparency, ethical data usage, inclusivity, and continuous improvement through learning.

A risk-based approach to regulating AI systems by categorizing and regulating them by their potential harm can foster innovation while addressing ethical concerns. Jobin et al. (2021) conducted a global review of AI ethics guidelines, finding that while there is some convergence on principles like transparency and justice, there remains significant variation in how these principles are interpreted and implemented. While AI ethics guidelines are a crucial step toward responsible development and use of AI, they are not without flaws. These guidelines often suffer from a lack of clarity, practical applicability, and enforceability. Implementing AI should align with the purpose and vision of the organization.

1.5 Innovative Teaching Tools and Teacher Support

AI enhances learning experiences by driving the development of innovative teaching tools through streamlining administrative tasks. AI-powered Virtual Reality (VR) and Augmented Reality (AR) applications create immersive and engaging learning environments, allowing students to explore historical events, dissect virtual organs, or practice complex skills in a safe and controlled environment. AI-powered ITS provides personalized instruction, feedback, and practice exercises, adapting to individual student needs and learning styles. (Anderson et al., 2008). AI can enhance content delivery with visual aids, making classroom management more effective.

A study by the University of Oxford found that students significantly improved their knowledge retention and problem-solving skills who used VR for learning. An ISTE study in 2019 confirmed that 88% of teachers believe AI is used to personalize learning and improve student outcomes. Similarly, a NEA report of 2020 found that 75% of teachers believe AI-powered tools improve their teaching effectiveness.

Chatbots and virtual assistants powered by AI can answer questions from students, offer assistance, and even automate administrative tasks, allowing teachers more time for more individualized interactions with students.

Computer-based intelligence can democratize training by making learning assets more open. With an estimated 770 million people who are not literate and 244 million children who are not enrolled in school worldwide, AI presents unique opportunities to make access to marginalized groups easier. Technology-driven educational tools are affordable and can close the educational access and quality gaps.

1.6 Challenges and Ethical Considerations

AI also presents challenges, along with numerous benefits. Only 10% of educational institutions surveyed by UNESCO have developed policies on generative AI, highlighting the need for robust policy frameworks (UNESCO, 2023). Moreover, there is a risk of perpetuating existing biases, particularly for children of colour who already face inequality around technological advancements.

By providing personalized learning experiences, AI is changing education**,** enhancing teaching capabilities, expanding access to education, and introducing innovative learning tools. However, successful integration requires careful consideration of ethical implications and the development of appropriate policies to maximize benefits while mitigating risks.

It will be crucial to shape the future of education to strike a balance between the promise of AI and its responsible application as we navigate this transformative era. The transformative influence of AI on the education sector is undeniable. By personalizing learning, analyzing data, and developing new teaching tools, all students will benefit from an equitable, engaging, and efficient learning environment provided by AI. As AI technology continues to evolve, its impact on education will grow, shaping the future of learning and preparing students for a world increasingly driven by data and automation.

By automating administrative tasks, AI tools enable educators to concentrate more on student interaction and individualized instruction. This synergy between AI and human educators creates a more unique and powerful learning climate.

This article explores the evolving landscape of AI in education, examining its potential to personalize learning, analyze data, and develop innovative teaching tools. It also delves into the exciting potential of AI in education, exploring its multifaceted benefits while acknowledging the critical need for responsible and ethical implementation.

2. Theoretical Frameworks for AI in Education

Integrating Artificial Intelligence (AI) with education necessitates a multifaceted theoretical approach covering technological, pedagogical, cognitive, and ethical dimensions. This framework synthesizes several theories and models to provide a comprehensive understanding of implementing AI effectively and responsibly in educational settings.

The Technological Pedagogical Content Knowledge (TPACK) framework provides a valuable model for integrating technology into education. By considering the dynamic interplay between content, pedagogy, and technology, Educators can create more effective and engaging learning environments. Mishra and Koehler introduced the TPACK framework in 2006, which offers a valuable lens for understanding the integration of AI in education and adds the technology dimension to Lee Shulman's Pedagogical Content Knowledge (PCK) concept. It emphasizes that content knowledge requires thoughtful pedagogical uses of technology in teaching.

2.1 Components of TPACK

- Content Knowledge (CK): Subject matter understood by the teacher, including knowledge of concepts, theories, ideas, organizational frameworks, and established practices within a particular discipline.
- Pedagogical Knowledge (PK): This involves the Teaching and Learning methods and processes. It encompasses understanding how students learn, classroom management, lesson planning, and assessment.
- Technological Knowledge (TK): This includes knowledge about various technologies, ranging from traditional tools like chalkboards to advanced digital tools like educational software and online resources.

The relevance of the TPACK framework in the context of Artificial Intelligence (AI) in education is significant. Integrating AI into the TPACK framework involves understanding how AI can enhance each of the knowledge domains:

- AI and Content Knowledge (CK): AI can provide adaptive learning platforms that tailor content to the individual needs of students
- AI and Pedagogical Knowledge (PK): AI can offer insights into student learning patterns and outcomes, helping teachers refine their pedagogical strategies by identifying the most suitable teaching methods for different sets of learners.
- AI and Technological Knowledge (TK): Teachers need to understand how to use AI tools effectively by knowing AI algorithms, data privacy issues, and the ethical implications of AI in education.

In the context of AI, the TPACK framework includes AI-specific knowledge, forming what some researchers call the Intelligent-TPACK framework. It emphasizes the importance of understanding AI tools and their capabilities for teaching effectively.

The Intelligent-TPACK framework highlights the importance of

1. Technological Knowledge (TK): Understanding AI tools and their capabilities
2. Pedagogical Knowledge (PK): Knowing how to teach with AI effectively
3. Content Knowledge (CK): Mastery of subject matter
4. AI-Pedagogical Knowledge: Understanding how AI can enhance teaching methods
5. AI-Content Knowledge: Recognizing how AI can represent and manipulate subject matter

Including AI within this framework further enhances its relevance, offering new possibilities for personalized and data-driven education. The AI-TPACK framework highlights the need for educators to develop competencies in AI-related content knowledge, AI-enhanced pedagogical approaches, and the technical skills required to implement AI tools in educational settings.

Considering the dynamic interplay between content, pedagogy, and technology, educators can create more effective and engaging learning environments. By integrating AI into the TPACK framework, educators can develop a more nuanced approach to teaching that leverages the strengths of AI while maintaining a robust pedagogical foundation. This integration can lead to more effective and personalized learning.

Connectivism: Complementing the TPACK framework, Connectivism, proposed by Siemens in 2005) offers insights into learning in the digital age. Connectivism posits the distribution of knowledge across networks of connections and that learning involves the ability to construct and traverse these networks.

This theory is particularly relevant to AI in education for it emphasizes the importance of navigating and making sense of information in rapidly changing digital landscapes. In the context of AI, connectivism supports the idea of AI as a tool for creating and navigating complex information networks, facilitating personalized learning pathways, and enabling continuous learning. Recent applications of connectivism in educational research have demonstrated its relevance in understanding how learners interact with AI-enhanced learning environments and how these interactions can foster self-directed learning and knowledge creation.

Adaptive Learning Theory, which draws from Cognitive Science and Computer Science, provides a framework for understanding how AI can personalize education. Learning is effective when instruction aligns with the individual needs, preferences, and abilities. AI-powered adaptive learning systems can assess learner knowledge and skills in real-time, provide personalized content and pacing, offer immediate feedback and remediation, and adapt to learner progress. Research has shown that adaptive learning systems can significantly improve learning outcomes and student engagement. The application of adaptive learning theory in AI-enhanced education aligns with the broader goal of creating more effective and personalized learning experiences.

The Cognitive Load Theory (CLT) developed by Sweller offers valuable insights into using AI to manage cognitive load in learning processes. CLT posits that learning is most effective when the cognitive load optimizes working memory. In the context of AI in education, this theory can guide the design of AI-powered learning environments that help manage intrinsic, extraneous, and germane cognitive load. For instance, AI tools break down complex tasks, provide just-in-time information, and offer scaffolding to support learners, reducing extraneous cognitive load and allowing more cognitive resources for learning.

Self-Determination Theory (SDT), proposed by Ryan and Deci in 2000, provides a framework for understanding motivation in learning contexts. SDT emphasizes the importance of satisfying three basic psychological needs: autonomy, competence, and relatedness. In the context of AI in education, SDT can guide the design of AI-enhanced learning environments that support these needs. For example, AI systems can provide personalized learning paths (supporting autonomy), adaptive challenges (supporting competence), and collaborative learning opportunities (supporting relatedness). However, it is crucial to consider how AI implementation potentially undermines these needs by over-prescribing learning paths or reducing human interaction.

The Zone of Proximal Development (ZPD) concept introduced by Vygotsky in 1978 describes the difference between what a learner can do without help and what they can do with guidance and encouragement. This concept is particularly relevant to AI in education, as AI systems can potentially provide the scaffolding and support needed to help learners progress through their ZPD. AI-powered tutoring systems, for instance, can adapt their level of support based on current abilities, gradually reducing assistance as the learner becomes more proficient.

The Capability Approach, developed by Sen, offers a framework for considering the broader impacts of AI in education on individual freedoms and societal development. This approach emphasizes the importance of expanding an individuals' capabilities – their freedom to achieve various functions. In the context of AI in education, the Capability Approach can guide how technologies might expand or limit learner capabilities and freedoms. It prompts us to consider the immediate learning outcomes and the long-term impacts of pursuing their goals and participating fully in society.

The Theory of Multiple Intelligences proposed by Gardner suggests that individuals possess different types of intelligence, including linguistic, logical-mathematical, spatial, musical, bodily kinaesthetic, interpersonal, intrapersonal, and naturalistic. This theory has implications for using AI to cater to diverse learning styles and intelligences. AI systems can potentially identify and adapt a learner's strengths across these different intelligences and provide personalized learning experiences that leverage their unique cognitive profiles.

The Learning Analytics framework, proposed by Siemens and Long in 2011, provides a structure for understanding how data about learners and learning environments can be collected, analyzed, and used to

improve education. In the context of AI, learning analytics enhanced through machine learning algorithms can process vast amounts of data to identify patterns, predict outcomes, and provide actionable insights. This framework is crucial for understanding AI to monitor and support student progress, inform instructional decisions, and improve educational systems at various levels (Siemens and Long, 2011).

The Knowledge Discovery in Databases (KDD) process model, introduced by Fayyad in 1996, offers a systematic approach to extracting knowledge from data. In the context of AI in education, the KDD process can guide the development and implementation of educational data mining techniques. These techniques can uncover hidden patterns in educational data, predict student performance, and provide personalized recommendations for learning.

The Diffusion of Innovations theory, proposed by Rogers in 1962, provides a framework for understanding how new ideas and technologies spread through social systems. This theory is relevant to the adoption of AI in education, as it can help explain and predict the rate and pattern of AI adoption among educators and educational institutions. Understanding the factors that influence the adoption of AI technologies can inform strategies for successful implementation in academic settings.

Ethical AI Framework, proposed by researchers like Luckin in 2016 emphasizes the need for transparency, accountability, privacy, and fairness in AI-enhanced educational settings. Holmes further developed this framework in 2021, proposing ethical guidelines for AI in education to include consideration of human agency, well-being, and social justice. These ethical considerations are critical given the potential for AI to exacerbate existing educational inequalities if not implemented thoughtfully.

Integrating these theories and frameworks provides a comprehensive approach to understanding and implementing AI in education. The TPACK and AI-TPACK frameworks guide the development of teacher competencies in integrating AI into their practice. Connectivism offers insights into how learning occurs in networked digital environments. Adaptive Learning Theory and Cognitive Load Theory inform the design of personalized, efficient learning experiences. Self-determination theory and the Zone of Proximal Development concept guide the creation of motivating and supportive learning environments.

The Capability Approach and Multiple Intelligences Theory ensure that AI implementation considers diverse learner needs and broader societal impacts. Learning Analytics and the KDD process model provide frameworks for leveraging educational data. The Diffusion of Innovations theory helps in adopting AI technologies in education. Finally, ethical frameworks ensure the implementation of AI responsibly and equitably.

However, challenges remain in fully realizing the potential of AI in education while adhering to these theoretical principles. Future research should focus on developing practical guidelines for implementing these frameworks in diverse educational contexts and evaluating their long-term impact on learning outcomes and educational equity. As AI evolves, these theoretical frameworks will require revision and adaptation to address new opportunities and challenges in AI-enhanced education.

3. Personalized Learning with AI

A statistical report from 2023 indicates that AI will upgrade 47% of learning management systems within the next three years. This trend suggests that AI is becoming more popular among the learners. Educational institutions must focus on personalized learning for learners by understanding their varied needs, abilities, and concerns. Personalized learning aims to be inclusive, benefiting all students rather than focusing on specific groups. According to Claned's 2024 report, personalized learning offers several potential benefits:

- **Adaptive Learning**: A report by Knewton in 2021 showed a 62% increase in test scores among students who used AI-powered adaptive learning programs in the USA. Adaptive learning allows educators to determine what content to teach and lets students skip material they already know, enabling them to focus on their personalized learning plans.
- **Targeted Interventions**: AI can help educators identify and address student difficulties. With AI, teachers can provide practice exercises, remedial classes, and tailored materials to help students overcome challenges. Organizations like Dream Box analyze student performance and intervene in their learning process by providing learning materials based on their capacity to understand the concepts.
- **Dynamic Content Delivery**: AI assists in developing personalized content for an individual student that suits the expectations and learning patterns. Such contents are delivered through various formats like videos, interactive sessions, and simulation activities, enhancing the learning experience.

- **Enhanced Accessibility**: The role of AI in making education accessible to all students is incredible. It provides content tailored to the needs of students with speech disabilities and slow learners, ensuring they can access and engage with educational materials effectively.

4. Data Analysis and Insights

Data analysis and statistics are crucial aspects of an AI-based education system. For instance, Data analysis and statistics in the education industry are supported by learning analytics to give more insight into student performance. Learning analytics involves systematically collecting and reporting data based on learner activities. AI enhances learning analytics by automating data analysis, providing insights into trends and patterns in student performance, predicting outcomes, and identifying learning gaps.

Educational data analysis transforms raw data into meaningful information for educators, administrators, and policymakers, creating an environment conducive to excellence. Data analysis will help educators for early intervention by understanding the patterns in the student's academic performance.

Data analytics helps students understand their learning progress. Integrating AI with data analytics motivates students to learn creatively and develop their unique capabilities. These tools aid students in grasping complex concepts and encourage self-learning. For educators, AI and data analytics tools assist in customizing the curriculum to meet the unique needs of students. The predictive analysis comes under AI-driven data analysis, identifies patterns in historical data, and predicts future outcomes, such as grades, scores, and overall performance.

AI tools combined with data analysis enable educators to monitor student performance by tracking response times, identifying the types of questions students skip, and understanding preferred learning resources. Technologies like emotion cues help educators gauge student interaction and interest levels through facial expressions.

5. New Teaching Tools Enabled by AI

The role of an educator evolves with the dynamic nature of the educational sector, and recent advancements in AI have significantly contributed to this dynamism. AI tools can enhance creativity and make class management easier for educators. Some of the AI tools for educators are

- **AudioPen**: AudioPen converts voice into text, offering an efficient way for educators to manage their time while preparing book chapters, materials, or lesson plans.
- **Canva Magic Write**: The Magic Write feature of Canva uses AI to transform text into images, fostering creativity among educators as they prepare visually appealing and engaging presentations for their students.
- **Curipod**: Curipod enables educators to create customized learning materials quickly, allowing students to access topics tailored to their learning needs.
- **Eduaide AI**: This tool assists educators in lesson development by providing over 100 resources for generating high-quality content. It also offers translation into 15 languages, making it easier for slow learners to understand the material in their native language.
- **OpenAI (ChatGPT)**: Both educators and students use ChatGPT for their ability to provide prompt answers with detailed explanations and suggestions. While the original version has limitations in content quality, ChatGPT-4 offers more reliable information.
- **Quizizz**: Quizizz helps educators create quizzes tailored to student requirements, segregate difficult questions, check for grammatical errors, and design questions based on real-time scenarios.
- **Slidesgo**: Slidesgo simplifies presentation preparation by providing access to Google Slides. Educators only need to input a topic, and Slidesgo asks for the desired tone (casual, professional, etc.). Slidesgo then converts the instructions into an appealing presentation.

6. Accessibility and Inclusion

Data is everything nowadays. With the advent of AI, data, and learning have become accessible to everyone. For instance, AI can convert text-based content to spoken words, making it accessible to people with learning disabilities. It also aids students by providing subtitles and closed captioning to facilitate learning. AI can enhance the lives of people with learning disabilities. Seeing AI, an app developed by Microsoft, reads documents aloud to assist visually impaired individuals. Adaptive learning tools of AI adjust the difficulty level of educational content based on performance, thus providing a platform where all students can learn at their own pace.

Accessibility is a fundamental right, and AI-embedded tools should give everyone inclusivity in the learning process. To achieve a high level of inclusivity, teachers need to adopt various teaching techniques to meet the diverse needs of their students.

7. Lifelong Learning and Career Development

AI is revolutionizing every industry, including education, making it crucial for individuals to remain competitive in their respective fields through Lifelong learning. Lifelong learning extends beyond traditional education, encouraging proactive learning to sustain long-term career success. AI-driven education systems will enhance lifelong learning by creating flexible learning pathways, intelligent content delivery, and continuous access to learning materials and mentors.

Individuals looking to upgrade their knowledge and explore new skills increasingly rely on AI generative tools. These tools bridge the gap between traditional learning and real-world experiences, facilitating experiential learning. Career-driven learning helps individuals understand their strengths, interests, and aspirations, guiding them to diverse professional opportunities.

AI in education fosters continuous intellectual and professional growth, addressing the unique needs and goals of the student. It transforms the one-size-fits-all education model into customized learning, allowing learners to acquire skills at their own pace. Lifelong learning, enhanced by AI, supports upskilling and reskilling to meet industry expectations. As the job market evolves, industries expect candidates to possess critical thinking, creativity, and emotional intelligence. Lifelong learning helps individuals remain resilient and adaptable in their workplaces. AI-driven learning analytics enable lifelong learners to understand their trajectories, identify areas for improvement, set realistic goals, and measure their performance.

8. Conclusion

Artificial Intelligence is revolutionizing education, offering personalized learning experiences and improved outcomes. However, its implementation requires balancing between innovation and ethical considerations. Theoretical frameworks like AI-TPACK provide a robust foundation for integrating AI into educational settings.

While AI-driven adaptive learning platforms tailor content and interventions to individual needs, significantly boosting engagement and results. Advanced analytics enhance data-driven decision-making, enabling early interventions and informed educational strategies. Moreover, AI empowers educators with innovative teaching tools, from voice-to-text converters to immersive VR and AR experiences, fostering creativity and efficiency in the classroom.

AI technologies also enhance accessibility, supporting diverse learners and promoting equal opportunities. They facilitate lifelong learning by creating flexible pathways for continuous skill development, which is crucial in the current dynamic job market. However, the integration of AI in education faces challenges such as accessibility gaps, teacher preparedness, data privacy concerns, potential developmental impacts, high implementation costs, and risks of technology addiction.

AI integration demands a clear purpose, regulatory compliance, AI literacy among educators, a balanced risk-benefit analysis, academic integrity, and human agency. Continuous evaluation of AI systems can ensure they serve educational goals effectively. As we embrace AI's potential, prioritizing ethical implementation and human-centric approaches will be vital. By doing so, we can create an inclusive, effective, and personalized learning environment that prepares students for an AI-driven future while nurturing their unique human qualities.

References

Adlawan, D., 2024. The Pros and Cons of AI in Education and How It Will Impact Teachers in 2024.Classpoint.io. Available at: https://www.classpoint.io/blog/the-pros-and-cons-of-ai-in-education

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (2008). Cognitive tutors: Lessons learned. The Journal of the Learning Sciences, 21(1), 1-32.

Bill & Melinda Gates Foundation. (2019). Personalized Learning: The Next Generation of Education. Retrieved from https://www.gatesfoundation.org/ideas/articles/personalized-learning-the-next-generation-of-education

Biswas, G., Jeong, H., Kinnebrew, J.S., Sulcer, B. and Roscoe, R.D., 2015. Measuring Self-Regulated Learning Skills through Social Interactions in a Teachable Agent Environment. Research and Practice in Technology Enhanced Learning, 10(1), pp. 1-30.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in neural information processing systems*, *33*, pp.1877-1901.

Brusilovsky, P. (2001). Adaptive hypermedia. User modeling and user-adapted interaction, 11, 87-110.

Dignum, V., 2019. Responsible artificial intelligence: how to develop and use AI in a responsible way (Vol. 2156). Cham: Springer.

Floridi, L. and Cowls, J. (2021). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1), pp. 150-179.

Jobin, A., Ienca, M. and Vayena, E. (2021). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), pp. 389-399.

Koedinger, K. R., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom.

Siemens, G. and Long, P., 2011. Penetrating the fog: Analytics in learning and education. EDUCAUSE review, 46(5), p.30.

UNESCO, 2023. Generative Artificial Intelligence in education: Think piece by Stefania Giannini. Available at: https://www.unesco.org/en/articles/generative-artificial-intelligence-education-think-piece-stefania-giannini

Effective Continuous Quantitative Measures for End-to-End AI Guardrails

Jason Pittman
Booz Allen Hamilton, McLean, USA

pittman_jason@bah.com

Abstract: Large Language Models such as ChatGPT have brought cutting-edge AI systems into the cultural zeitgeist. As a result, AI is no longer an isolated fief of academia or forward-leaning businesses. There are more than 35 million visits to open-source models in public repositories monthly. Clearly, the general technology community has caught onto the power of such systems and is keen to harness the promise of efficiency, productivity, and enhanced capability. Concurrent to this uptrend, AI systems are understood to be potentially vulnerable to various ethical issues. Such issues range from bias and fairness, to explainability and trustworthiness. More than mere theory, such vulnerabilities have manifested in mainstream settings such as politics, medicine, and law. The ethical implementation and operation of AI systems is, therefore, of critical interest as the democratization of such systems gains accelerates. However, there is an ongoing challenge insofar as there is little consensus on what constitutes quantitative ethical and responsible AI guardrails. This leaves AI practitioners without sufficient guidance to implement systems reasonably free from societal level harm. Accordingly, this work presents a structured taxonomy and concept matrix consisting of 39 discrete guardrails arrayed across a three-phased AI system lifecycle. Measure families further organize measures in terms such as bias mitigation, adversarial robustness, and anomaly monitoring. Then, I provide specific quantitative metrics for each measure construct. The intended takeaway is for AI practitioners to have the means to select appropriate and effective metrics for assuring ethical and responsible guardrails.

Keywords: Artificial intelligence, AI governance, Responsible AI, AI guardrails, Quantitative controls

1. Introduction

We are firmly in the age of artificial intelligence (AI). The technology now permeates decision-making in a plethora of fields ranging from healthcare to finance, and from transportation to entertainment. Moreover, generative AI- ChatGPT as a prime example- has become a popular culture icon and mainstream technology. Thus, as AI systems become more pervasive, ensuring such system operate ethically and responsibly is paramount (Cheng, Varshney, & Liu 2021). This need has given rise to the concept of *responsible AI*, which emphasizes the development and deployment of AI systems that are fair, transparent, and aligned with societal values (Mitchell et al 2019).

Responsible AI involves a comprehensive approach to the entire lifecycle of AI systems. The lifecycle can be conceptualized initial design, training, and deployment. Further, development can be divided into three phases: deploy, maintain, and continuous workflow. Throughout each phase, Responsible AI aims to assure AI systems do not cause harm and operate within responsible, ethical parameters. In this sense, a critical component of responsible AI is the implementation of *controls* during the training phase. Controls refer to measures and techniques applied to the training data and algorithms to mitigate biases, enhance fairness, and ensure transparency in decision-making processes (Mehrabi et al 2021).

However, the responsibility of ensuring ethical AI does not end once the system is deployed. AI systems must continuously adapt to new data and evolving real-world conditions. This is where Responsible AI *guardrails* come into play. Guardrails are mechanisms and practices implemented in production environments to monitor, guide, and control the behaviour of AI systems in real-time (Raji & Buolamwini, 2019). Thus, guardrails are the sole means for assuring responsible and ethical standards in operational AI systems. Therein lies the problem.

Developing guardrails that scale across different applications and contexts is challenging. What works for one application may not be suitable for another, necessitating context-specific solutions (Holstein et al 2019). Further, there is little consensus on what constitutes quantitative ethical and responsible AI guardrails. Further, there is a lack of standardization in the methods and metrics used to implement and evaluate guardrails (Raji & Buolamwini, 2019). This creates great difficulty for AI engineers attempting to leverage responsible AI principles.

Accordingly, the purpose of this work was to demonstrate a structured taxonomy and concept matrix consisting of 38 discrete controls arrayed across a three-phased AI system lifecycle. The taxonomy and concept matrix extend the framework proposed by Li et al (2023) by isolating controls capable of acting as guardrails. Also, this work is not limited by those principles contributed to trustworthy AI. Meaning, our aim is to provide

AI practitioners with the means to select appropriate and effective guardrails for assuring end-to-end ethical and responsible AI.

The remainder of this study consists of four sections. The Related Work details the background research contributing to an understanding of responsible AI as a field as well as defining, comparing, and contrasting controls and guardrails. The Method describes the mechanisms used to construct the taxonomy and concept matrix. Naturally, the Findings demonstrate the output of the Method. Lastly, recommendations and ideas for future work appear in the Conclusions.

2. Related Work

The existing work related to this study includes an overview of responsible AI in general, along with a discussion on controls and guardrails. The review is not intended to be exhaustive, however. Instead, this section endeavors to develop sufficient understanding of responsible AI controls and guardrails such that you, the reader, may appropriately synthesize the findings and conclusions.

2.1 Responsible AI

The concept of AI is common knowledge. Much less common is how ethical and responsible AI can be achieved. To start, an operational definition of responsible AI typically encompasses practices ensuring AI systems are developed and deployed in ways that are ethical (Floridi et al 2019; Mittelstadt et al 2016). Ethical, in this context, generally implies principles such as fairness, transparency, privacy, security, and trustworthiness.

The field of responsible AI has advanced rapidly over the past decade. In a short time, researchers have developed robust theoretical and applied conceptual frameworks. In part, the rapid expansion has been driven by notable examples of harm resulting from a lack of responsible AI. Such examples include discriminatory sentencing and parole decisions in the U.S. justice system (Angwin et al 2022) and Amazon's recruitment tool becoming biased against women (Dastin 2022). Another part of the expansion is increasing legal and regulatory requirements such as U.S. President Biden's Executive Order and the EU's AI Act (Wörsdörfer 2024).

Fundamentally, an AI system can be considered ethical and responsible when proper controls and guardrails are in place. In short, these are the technical mechanisms used to assess or implement ethical and responsible AI principles. While controls and guardrails are related, the two are not identical constructs. Thus, we should take some time to discuss both and how each interacts with AI.

2.2 Controls and Guardrails

Controls refer to measures and techniques applied during the training phase of AI development to ensure that the AI system behaves ethically and responsibly (Mitchell et al 2019; Mehrabi et al 2021). It is critical to note controls are associated with data. The goal for controls is to address potential issues such as biases, fairness, explainability, and robustness before the system is deployed (Doshi-Velez & Kim, 2017). While not an exhaustive set, these issues are perhaps the most commonly addressed in the literature.

Definitionally, bias controls reduce or eliminate bias in training data or model outputs (Mehrabi et al 2021). Holstein et al (2019) suggested fairness controls ensure sensitive elements in the data are treated with parity.

Explainability controls make the model's decisions more understandable to humans (Doshi-Velez & Kim, 2017). Robustness controls ensure the model performs well under various conditions and is resilient to adversarial attacks (Varshney & Alemzadeh, 2017).

Guardrails are mechanisms implemented in production systems to assure the behavior of AI systems in real-time (Raji & Buolamwini, 2019; Varshney & Alemzadeh, 2017). The aim is to ensure the AI system continues to operate responsibly and ethically during the system's deployment lifetime (Holstein et al., 2019). Perhaps confusingly, guardrails are in effect controls. In many instances, the name and purpose of a guardrail (e.g., fairness) is identical to the same applied during training of the AI. However, the guardrail is either embedded in the system or applied from an external source. Recall, controls in contrast are applied to data.

Make no mistake about it, an AI system not intentionally implemented and maintained with guardrails will operate ethically and responsibly only by chance. This is true even if controls were applied during training because the AI system will experience perturbations because of interacting with users, other systems, and itself during the system's deployed lifetime (Varshney & Alemzadeh, 2017).

Thus, guardrails are important to assure ongoing desirable behavior. This means anomalous behavior, ethical concerns, or unintended consequences are identified and addressed, thereby preventing harm (Raji & Buolamwini, 2019). Further, guardrails enable ongoing assessment and adjustment to ensure fairness is maintained and new biases are promptly identified and mitigated (Holstein et al 2019). Meanwhile, transparency and explainability are key to perception of an AI system's trustworthiness (Mitchell et al 2019). Of course, AI systems must be robust and safe as well. Here, guardrails enhance the robustness and safety of AI systems by allowing for continuous evaluation and resilience against adversarial interactions (Varshney & Alemzadeh, 2017).

All that stated, there are a variety of important open challenges in the field of responsible AI. Principally, while substantial tools and frameworks such as IBM's AI Fairness 360 (Bellamy et al 2018) exist to aid in the implementation of responsible AI controls, there is a notable gap in comprehensive, operational guidance for guardrails (Varshney & Alemzedah 2022). Addressing this gap requires developing construct-specific guidelines based on the AI system lifecycle and approach. I present just that in the remainder of this paper.

3. Method

A single research question motivated this work: *how can AI practitioners select appropriate and effective controls for assuring ethical guardrails in the systems they develop*. Selecting appropriate and effective guardrails assumes a set of such guardrails exists from which targeted subsets may be created. On one hand, the set does exist as represented by the corpus of research. On the other hand, however, the same corpus lacks organized, clear guidance on which guardrail or guardrails are best applied to an AI system relative to specific threats.

To this end, I used a meta-synthesis design to construct a taxonomy and concept matrix. A meta-synthesis was the most appropriate method to achieve this purpose as the method uniquely allows the generation of models, taxonomies, and theory from qualitative inputs (Barnett-Page & Thomas 2009; Thomas & Harden 2008). A reasonable alternative method could have been a systematic literature review had the input been quantitative.

Taking inspiration from the framework proposed by Li et al. (2023), I first reviewed the relevant literature indicated in Table 1 in that study. My review initially separated research into controls and guardrails based on the references from the source table in Li et al. (2023). Then, I extracted those references directly associated with guardrails while preserving the coupled lifecycles and approaches. Finally, I reached back into the literature to locate sources for seminal qualitative descriptions of the quantitative guardrail metrics.

4. Findings

The findings are organized according to a synthesis of the AI system lifecycles demonstrated by Li et al. (2023). Fundamentally, there are two conceptual expressions necessary to demonstrate our findings such that the findings are actionable for practitioners. We present each expression nested underneath the parent Li et al. lifecycle. I opted to prefix the lifecycle phase with an approach to construct the foundation for a narrative inquiry. Further structure is imparted by two additional layers: measure families and measures. Finally, we attach a seminal study demonstrating the specific metric in an applied quantitative context.

4.1 Assure-Deployment

The deployment of an AI system is the first phase of the system's lifecycle wherein guardrails can be implemented. Assuring an AI system deployment requires at least four measure families be implemented (Figure 1). The measure families are independent from one another but contribute equally to the ethical and responsible state of the system.

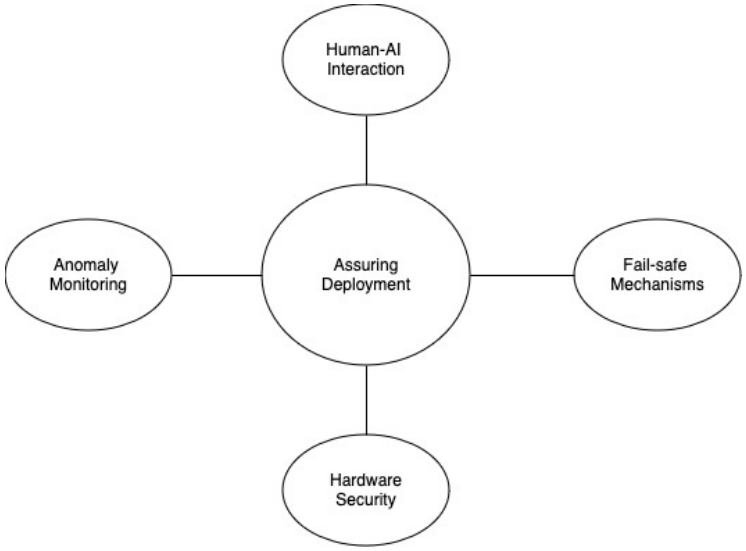


Figure 1: A taxonomy for assuring deployment of an AI system

The taxonomy serves as a high-level architecture for technical requirements. In turn, each measure family encompasses a set of one-to-one mapped measures and metrics (Table 1). One would expect to implement as many metrics as appropriate for the specific type of AI system under consideration.

Table 1: Concept matrix for assuring deployment of an AI system with guardrails

Assure - Deployment		
Measure Family	**Measures**	**Metric**
Anomaly Monitoring	Performance Drift	Bayram et al 2022
	Data Drift	Gama et al 2014
	Model Confidence	Hu & Mordohai 2010
	Confusion Matrices	Gorodkin 2004
	Outlier Detection	Schölkopf et al 2001
	Temporal Change	Huang & Paul 2018
Human-AI Interaction	Trust Metrics	Ueno et al 2022
	Explainability Score	Lundberg & Lee 2017
	Autonomy and Control Measures	Goodrich et al 2001
Fail-safe Mechanisms	Ethical Compliance	Mökander & Floridi 2023
	Error Detection Rate	Gal & Ghahramani 2016
	Redundancy Effectiveness	Lakshminarayanan et al 2017
Hardware Security	Physical Intrusion Detection Rate	Garcia 2008
	Hardware Failure Rate	Huang et al 2005
	Tamper-Evident Effectiveness	Suh et al 2003
	Data Wipe Success Rate	Gutmann 1996

4.2 Auditing-Management

Having the means to audit an AI system once the system is in production is the essence of guardrails. Meaning, keep the pulse of the system is necessary to demonstrate whether the system is in an ethical, responsible state. This phase of the lifecycle encompasses seven measure families. It is worth mentioning some research suggests *explainability* and *transparency* are predicated on the inclusion of the other six measure families. For our purposes, we continue to assume all measure families are independent.

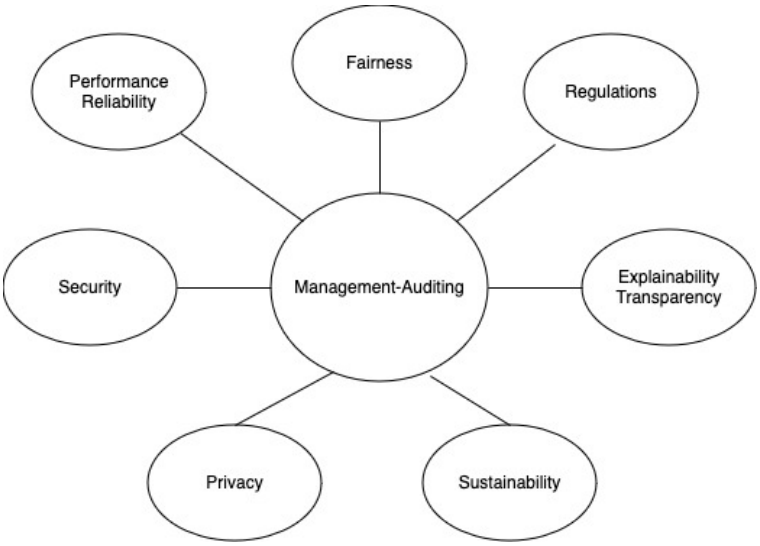


Figure 2: A taxonomy for auditing management of an AI system

The seven measure families provide technical structure for 15 discrete measures (Table 2). Again, operationally one would implement at least one metric from each measure family but also as many as might be appropriate given the type of AI system. If responsible AI controls were considered during training, three measure families- fairness, performance, and sustainability- would be directly transposable to the guardrail context.

Table 2: Concept matrix for auditing management of an AI system with guardrails

Auditing - Management		
Measure Family	**Measures**	**Metric**
Fairness	Disparate Impact Score	Feldman et al 2015
	Demographic Parity	Dwork 2012
Explainability & Transparency	Feature Importance Metrics	Hooker et al 2019
	Explanation Fidelity	Yeh et al 2019
Privacy	Differential Privacy	Abadi et al 2016
	Data Anonymization	Sweeney 2002
Security	Adversarial Resilience	Carlini & Wagner 2017
	Error Rate	Papernot et al 2016
	Anomaly Detection	Ali & Casola 2019
Performance	Accuracy, Precision, Recall, F1	Kim 2020
	System Availability	Kohli et al 2019
Regulations	Ethics Alignment	Gabriel 2020
Sustainability	Energy Efficiency	Danish 2023
	Carbon Footprint	Henderson et al 2020

4.3 Operations-Workflow

The operations workflow, or MLOPs, is a continuous pipeline of updates and changes to the underlying AI system. In other words, the measure families and metrics encompassed in this lifecycle phase are the means to assure the system *continues* to operate ethically and responsibly.

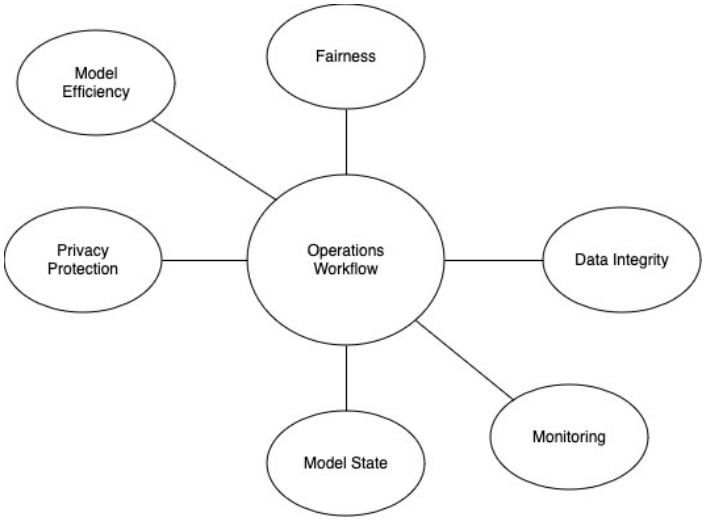


Figure 3: A taxonomy for operations workflow in AI systems

Two of five measure families- fairness and model efficiency- potentially carry over from both the training responsible AI controls as well as from the auditing-management phase. The other three represent measure families unique to any AI system operations workflow (Table 3).

Table 3: Concept matrix for operations-workflow phase with guardrails

Operations - Workflow		
Measure Family	**Measures**	**Metric**
Fairness	Disparate Impact Score	Breck et al 2017
	Demographic Parity	Zhang et al 2022
Data Integrity	Data Quality	Breck et al 2019
	Data Provenance Tracing	Moreau et al 2011
Model State	Model Robustness Score	Gheiros eta al 2019
	Model Reliability Score	Sculley et al 2015
Model Efficiency	Accuracy, Precision, Recall, F1	Amershi et al 2019
Monitoring	Anomaly Model Behaviour	Marzano et al 2022
	Model Drift Metrics	Lu et al 2018

5. Conclusions

AI is squarely planted in the cultural zeitgeist. As a result, AI is no longer an isolated fief of academia or science fiction. There are more than 35 million visits to open-source models in public repositories monthly. Clearly, the general technology community has caught onto the power of such systems and is keen to harness the promise of efficiency, productivity, and enhanced capability. Concurrent to this uptrend, AI systems are understood to be potentially vulnerable to various ethical issues. Such issues range from bias and fairness, to explainability and trustworthiness.

On one hand, existing research demonstrates a meaningful grasp of ethical and responsible AI *controls*. Further, to a more limited degree, the literature reveals a foundation for ethical and responsible AI *guardrails*. On the other hand, the same literature and research are bereft of practical guidance for implementing end-to-end guardrails. Thus, the goal of this work was to develop sets of taxonomies and guardrail matrices for practitioners to leverage when selecting appropriate responsible AI controls to implement.

Building on the foundation of AI system lifecycles and approaches from Li et al (2023) The findings demonstrate a praxis for AI practitioners to select appropriate and effective guardrails. My hope is for AI practitioners to leverage the taxonomies and guardrail matrices to implement responsible AI. This is achievable through simple narrative inquiry. For example- referencing Figure 1 and Table 1- one might inquire *what guardrail metric is most appropriate if I want to assure the performance drift of a deployed an AI system as*

part of my anomaly monitoring. The answer, given the findings, is one of the metrics assessed by Bayram et al (2022), specifically in Table 3.

One limitation therein is we lack a complete mapping of metrics onto AI system permutations. Thus, the indicated metric may not be applicable to the actual AI system given its construction or base model. In other words, the generalizability of existing metrics needs to be understood. However, when viewed as a starting place rather than a destination, we suggest the AI practitioner can extract the necessary guidance to satisfactorily answer such inquiries. Another limitation of this work is the lack of consideration for cutting edge and future types of AI systems. Generative AI, multi-modal, and neuro-symbolic architectures need applied responsible AI solutions, but existing metrics may not be applicable.

It is one thing to have a taxonomy or guardrail concept matrix outlining what can be measured in terms of responsible AI. It is something else entirely to have validated, functional examples to follow. For that reason, future constructive work is necessary to demonstrate proper guardrail implementations and associated effectiveness. Here, constructive research would be of immense benefit to the field and practitioners. Along such lines, future work is necessary to develop a complete mapping of existing metrics to existing types of AI systems. Doing so necessarily must include future leaning ideas covering frontier AI technology such as generative AI and neuro-symbolic AI. Failing to consider such futures now will only leave practitioners behind the responsible AI curve as generative and neuro-symbolic take over from traditional AI systems.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., & Zhang, L. 2016, "Deep Learning with Differential Privacy", *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308-318.

Ali, D. & Casola, L. eds. 2019, *Robust Machine Learning Algorithms and Systems for Detection and Mitigation of Adversarial Attacks and Anomalies: Proceedings of a Workshop*. National Academies Press.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., & Zimmermann, T. 2019, "Software Engineering for Machine Learning: A Case Study," *IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pp. 291-300.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L., 2022. “Machine bias”, In *Ethics of data and analytics*, pp. 254-264. Auerbach Publications.

Bayram, F., Ahmed, B.S. & Kassler, A. 2022, “From concept drift to model degradation: An overview on performance-aware drift detectors”, *Knowledge-based Systems*, 245, pp. 1-19.

Barnett-Page, E., & Thomas, J. 2009, "Methods for the Synthesis of Qualitative Research: A Critical Review." *BMC Medical Research Methodology,* Vol. 9, No. 59.

Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A. & Nagar, S. 2018, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias". *arXiv preprint arXiv:1810.01943*.

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. 2017, "The ML test score: A rubric for ML production readiness and technical debt reduction," *IEEE International Conference on Big Data (Big Data)*, pp. 1123-1132.

Breck, E., Zinkevich, M., Polyzotis, N., Whang, S.E., & Roy, S. 2019, "Data Validation for Machine Learning", *Proceedings of the 2nd SysML Conference*, pp. 1-11.

Carlini, N. & Wagner, D. 2017, "Towards Evaluating the Robustness of Neural Networks", *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39-57.

Cheng, L., Varshney, K.R. & Liu, H. 2021, “Socially responsible ai algorithms: Issues, purposes, and challenges”, *Journal of Artificial Intelligence Research*, Vol. 71, pp.1137-1181.

Danish, M.S.S. 2023, "AI and Expert Insights for Sustainable Energy Future", *Energies*, Vol. 16, No. 8, 3309.

Dastin, J. 2022, “Amazon scraps secret AI recruiting tool that showed bias against women”, In *Ethics of data and analytics*, pp. 296-299. Auerbach Publications.

Doshi-Velez, F., & Kim, B. 2017, “Towards a rigorous science of interpretable machine learning”, *arXiv preprint arXiv:1702.08608*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. 2012, "Fairness through Awareness", *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214-226.

Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. 2015, "Certifying and removing disparate impact", *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259-268.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. & Schafer, B., 2018, “AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations”, *Minds and machines*, No. 28, pp.689-707.

Gabriel, I. 2020, "Artificial intelligence, values, and alignment", *Minds and Machines*, Vol. 30, pp. 411–437.

Gal, Y., & Ghahramani, Z. 2016, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1050-1059.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. 2014, "A survey on concept drift adaptation", *ACM Computing Surveys (CSUR)*, Vol. 46, No. 4, pp.1-37.

Garcia, M. L. 2008, "The Design and Evaluation of Physical Protection Systems", 2nd ed., Burlington, MA: Butterworth-Heinemann.

Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F.A. 2019, "A Comprehensive Study on Robustness of Image Classification Models: Benchmarking and Rethinking", *arXiv preprint arXiv:1902.10171*.

Goodrich, M.A., Olsen, D.R., Crandall, J.W., & Palmer, T.J. 2001, "Experiments in adjustable autonomy", *Proceedings of IJCAI Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents*, pp. 162-167.

Gorodkin, J. 2004, "Comparing two K-category assignments by a K-category correlation coefficient", *Computational Biology and Chemistry*, Vol. 28, No. 5, pp. 367-374.

Gutmann, P. 1996, "Secure Deletion of Data from Magnetic and Solid-State Memory", *Proceedings of the 6th USENIX Security Symposium*, pp. 77-89.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. 2020, "Towards the Carbon Footprint of Machine Learning", *arXiv preprint arXiv:1910.09700*.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. 2019, "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?", In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-16.

Hooker, S., Erhan, D., Kindermans, P.J., & Kim, B. 2019, "A Benchmark for Interpretability Methods in Deep Neural Networks", *arXiv preprint arXiv:1806.10758.*

Hu, X. & Mordohai, P. 2010, "Quantitative evaluation of confidence measures in a machine learning world", *International Conference on Computer Vision (ICCV)*, pp. 2353-2360.

Huang, B., Li, X., Li, M., Bernstein, J., & Smidts, C. 2005, "Study of the impact of hardware fault on software reliability", *Proceedings of the 16th IEEE International Symposium on Software Reliability Engineering (ISSRE)*, IEEE, pp. 10-72.

Huang, J. & Paul, M.J. 2018, "Examining Temporal Changes in a Language Model's Performance", *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1479-1489.

Kim, I. 2020, "TEVV (Test, Evaluation, Verification & Validation) Strategy for Ethical AI (Artificial Intelligence) Development", *Journal of The Korean Institute of Defense Technology*, Vol. 2, No. 1, pp.001-004

Kohli, A., Gupta, N., & Vashistha, A. 2019, "Techniques for ensuring high availability of AI systems", *International Journal of Computer Applications*, Vol. 182, No. 37, pp. 45-52.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2017, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles", *Advances in Neural Information Processing Systems (NIPS)*, pp. 6402-6413.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. 2018, "Learning under Concept Drift: A Review," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 31, No. 12, pp. 2346-2363.

Lundberg, S.M. & Lee, S.I. 2017, "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems (NIPS)*, pp. 4765-4774.

Marzano, A., Bellini, E., & Blasi, F. 2022, "Real-Time Anomaly Detection in MLOps Pipelines: A Comprehensive Approach", *Electronics*, Vol. 11, No. 11, p. 3640.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. 2021, "A Survey on Bias and Fairness in Machine Learning", *ACM Computing Surveys (CSUR)*, Vol. 54, No. 6, 1-35.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. & Gebru, T. 2019, "Model cards for model reporting.", In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229.

Mökander, J. & Floridi, L., 2023, Operationalising AI governance through ethics-based auditing: an industry case study. *AI and Ethics*, Vol. 3, No. 2, pp.451-468.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., & Myers, J. 2011, "The Open Provenance Model Core Specification (v1.1)", *Future Generation Computer Systems*, Vol. 27, No. 6, pp. 743-756.

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. 2016, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks", *IEEE Symposium on Security and Privacy (SP)*, pp. 582-597.

Raji, I. D., & Buolamwini, J. 2019, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products", In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 429-435.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. 2001, "Estimating the Support of a High-Dimensional Distribution", *Neural Computation*, Vol. 13, No. 7, pp. 1443-1471.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. 2015, "Hidden Technical Debt in Machine Learning Systems", *Advances in Neural Information Processing Systems (NIPS)*, pp. 2503-2511.

Suh, G.E., Clarke, D., Gassend, B., Van Dijk, M. & Devadas, S. 2003, "AEGIS: Architecture for tamper-evident and tamper-resistant processing", In *ACM International Conference on Supercomputing 25th Anniversary Volume*, pp. 357-368.

Sweeney, L. 2002, "k-Anonymity: A Model for Protecting Privacy", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 557-570.

Thomas, J., & Harden, A. 2008, "Methods for the Thematic Synthesis of Qualitative Research in Systematic Reviews." *BMC Medical Research Methodology,* Vol. 8, No. 45.

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. 2022, "Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods", *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, Article No. 254. Available at: https://doi.org/10.48550/arXiv.2205.00189.

Varshney, K. R., & Alemzadeh, H. 2017, "On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products", *Big Data*, Vol. 5, No. 3, 246-255.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K. & Cave, S., 2019, *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. London: Nuffield Foundation.

Wörsdörfer, M., 2024, "Biden's Executive Order on AI and the EU's AI Act: A Comparative Computer-Ethical Analysis", *Philosophy & Technology*, Vol. 37, No. 3, p.74.

Yeh, C.K., Hsieh, C.Y., Suggala, A., Inouye, D.I. & Ravikumar, P.K. 2019, "On the (in)fidelity and sensitivity of explanations", *Advances in neural information processing systems*, *32.*

Zhang, L., Ghassemi, M., Brugere, I., Dalmasso, N., Mishler, A., Potluru, V., Balch, T., & Veloso, M. 2022, "Conditional Demographic Parity Through Optimal Transport", *NeurIPS 2022*.

Comparative Analysis of YOLOv5/v8/v9 for Object Detection, Tracking, and Human Action Recognition in Combat Sports

Evan Quinn and Niall Corcoran
Technological University of the Shannon: Midlands Midwest, Limerick, Ireland

evan.quinn@tus.ie
niall.corcoran@tus.ie

Abstract: YOLO models are widely used object detectors in computer vision (CV). This study investigates the relative performance of YOLOv5, YOLOv8, and YOLOv9 for object detection, tracking, and human action recognition in combat sports. The models were evaluated using curated datasets encompassing various combat scenarios, athlete movements, and equipment configurations. Pre-processing protocols and augmentation techniques were applied to improve model accuracy and generalizability, including automated orientation correction, image dimension standardisation, contrast enhancement, and methods such as zoom, rotation, shear, and grayscale conversion. The key findings provide insight into the comparative performance of the models across various evaluation metrics, such as precision, recall, and mean average precision. Each model's ability to detect, track, and recognise human actions in dynamic combat sports environments is evaluated. Computational efficiency and real-time performance were assessed as these are important indicators for practical applications in coaching, training, and competitive scoring systems. The findings suggest that YOLOv8 offers the best balance of precision and recall, making it particularly suitable for real-time applications in combat sports analytics. This study contributes to advancing CV technologies in combat sports analytics, with potential implications for improving athletic training methods, facilitating personalised coaching interventions, and enhancing objectivity and consistency in competitive scoring processes in combat sports.

Keywords: YOLOv5, YOLOv8, YOLOv9, Object detection and tracking, Combat sports, Computer vision

1. Introduction

Recent advances in artificial intelligence (AI), computer vision (CV), and decision science are revolutionising many sectors, not least sports, with improvements in several areas including performance analysis, injury prevention, and viewer engagement (Kumar, 2022). The growing field of CV, fuelled by an abundance of visual data, has facilitated athletic performance monitoring and the automation of statistical analysis in sports (Host and Ivašić-Kos, 2022; Jaiswal *et al.*, 2020). For example, Deep Learning (DL) applied to CV has gained interest in sports video understanding, providing richer information to sports consumers and leagues (Van Zandycke *et al.*, 2022). These technologies could be especially effective in combat sports such as boxing, kickboxing, wrestling, and mixed martial arts (MMA), offering precise and unbiased evaluations that address the historical issue of human bias in scoring (e.g., Binter *et al.*, 2024). These disciplines are characterised by one-on-one engagements with specific rulesets and supervised by a referee. Expert judges are heavily relied on to identify athletes objectively, but this process is prone to human error. Furthermore, each combat sport requires a different strategy due to the varying complexities involved in each (Makovsky, 2020). For example, MMA scoring evaluates a wide range of techniques that may be more complex than those found in other combat sports, such as boxing, including strike energy and submission-hold skill (James *et al.*, 2017). New AI-driven technologies offer a fresh lens, promising precision, and unbiased evaluations.

CV has already shown promise in reducing human error in combat sports analysis, such as determining the velocity and impact of punches (Ishac and Eager, 2021) and measuring combat sports athlete performance using wearable sensors (Saponara, 2017). CV has the potential to provide several benefits for combat sports, including scientific training, more effective technical and tactical analysis, and the generation of performance metrics and statistics. Pang *et al.* (2022) highlight the versatility of CV in combat sports, claiming it promotes scientific training, strengthens technical and tactical scrutiny, and generates performance metrics. CV streamlines the analysis, providing trainers and athletes with data-driven insights for improving performance and tactics. Echeverria and Santos (2021) demonstrated this with a study on punch anticipation in karate, offering a comprehensive understanding of martial body movements. The contribution of DL to CV, demonstrated by Van Zandycke *et al.* (2022), further supports this approach. Models built from large datasets, with high-resolution images and detailed annotations, provide comprehensive insights for sports enthusiasts and professionals. According to Pang *et al.* (2022), direct comparison and similarity calculations can be used to evaluate human action attributes.

Using CV techniques in karate has resulted in significant advances in human action recognition (HAR) research for combat sports tracking (Echeverria and Santos, 2021). Convolutional neural networks (CNN) have proven

useful in image recognition and pixel data processing tasks in CV (Patil and Rane, 2021), with potential applications in combat sports. Recent advances in object detection and classification, such as the You Only Look Once (YOLO) architecture, have simplified the process, allowing for real-time performance with high accuracy (Yin *et al.*, 2020). According to Bochkovskiy *et al.* (2020), CNN combined with YOLO has paved the way for object detection and pose estimation in combat sports. HAR is useful for identifying and analysing human movement in videos, but it has been difficult to apply due to the large number of athletes participating in team sports such as football or volleyball (Host and Ivašić-Kos, 2022). Therefore, HAR may be more suitable for combat sports analysis due to the limited number of athletes in a competitive scenario. Accurate HAR is crucial in extracting meaningful information from combat sports videos. Pose detection, a widely used CV task, enables the prediction of human poses, facilitating HAR and action classification (Angelini *et al.*, 2019). Furthermore, the adoption of time-series analysis allows for accurate tracking of identifiers, including pose estimation features, addressing the challenges of long-range context analysis and vanishing gradient problems (Ning *et al.*, 2017). Several studies examine the different YOLO algorithms and compare them for performance and accuracy for real-time single-shot object detection and tracking (e.g., Atik *et al.*, 2022; Jiang *et al.*, 2022; Terven and Cordova-Esparza, 2023).

The primary goal of this study is to investigate the capability of CV for automating scoring in combat sports by comparing automated judgments to human evaluations. Given the intricacies of martial arts movements and the sport-specific rules, crafting a universally embraced scoring algorithm is challenging. It is important to determine the most appropriate architectures for performance and scoring analysis based on CV systems that can categorize athletes based on factors such as attire, stance, and physical characteristics. As part of the main research project, this paper sets out to report on a comparison of three versions of the YOLO model using the same dataset to determine which performs best on the computer hardware. This evaluation will assist with model selection for further research. The focus is to identify the most effective CV architectures for performance and scoring analysis, analyse combat sports footage, and categorise athletes. Additional goals are to determine effective data processing methods, establish best practices for curating labelled combat sports footage, and minimize inaccuracies from low-quality datasets.

1.1 YOLO Models

YOLO is a popular object detection model known for its speed and accuracy (Yin, Li and Fu, 2020). YOLO performs object detection as a regression problem and returns the class probabilities of the detected images. It is a real-time algorithm that detects various objects in images and videos. It was first introduced by Redmon *et al.* (2016) and has had several iterations since then. In this study, YOLOv5, YOLOv8, and YOLOv9 are selected to explore the advancements in the most recent versions and their applicability to combat sports analysis. YOLOv5, implemented in PyTorch and pre-trained on the COCO dataset (Veit *et al.*, 2016), serves as a robust baseline widely adopted for its balance between speed and accuracy Yang *et al.* (2020). It achieves real-time detection at over 140 frames per second (FPS) and has been successfully applied in various industrial applications. YOLOv8 introduces significant enhancements over previous versions, incorporating new elements such as attention modules, self-attention mechanisms, spatial pyramid pooling, and deformable convolutions, which improve detection accuracy and speed while reducing computational demands (Sohan *et al.*, 2024). These advancements make it particularly suitable for combat sports analysis, where precise and efficient detection is crucial. YOLOv9 incorporates new features such as Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN) to address data loss in deep networks, achieving superior results on the MS COCO dataset (Wang *et al.*, 2024). These innovations can potentially enhance detection accuracy in complex scenarios such as those found in combat sports. YOLOv6 and YOLOv7 are excluded from the analysis for specific reasons. YOLOv6, despite its optimisations for industrial applications, lacks widespread community support and open-source collaboration, making it less suitable for a comparative study aimed at general applicability. YOLOv7, while achieving impressive performance improvements and introducing real-time pose estimation and instance segmentation (Boesch, 2023), has been succeeded by YOLOv8, which builds upon YOLOv7's advancements and offers further enhancements with a more streamlined architecture.

Focusing on YOLOv5, YOLOv8, and YOLOv9 aims to assess the impact of the most recent developments in the YOLO family on object detection, tracking, and HAR in combat sports. This approach allows us to evaluate models representing significant milestones in the evolution of the YOLO architecture and determine their suitability. Some practical applications of YOLO include assembly-line object recognition (Liau and Ryu, 2021), real-time underwater maritime object detection (Yu *et al.*, 2022), facial mask detection (Yang *et al.*, 2020), and diabetic foot ulcer detection (Yap *et al.*, 2021). These examples underscore the effectiveness of YOLO architectures in

different domains, reinforcing the relevance of comparative analysis in combat sports. Table 1 summarises the main features and advancements of YOLO versions v5 to v9.

Table 1: Summary of YOLO Versions v5 to v9

Version	Year	Key Features	Advantages	Limitations
v5	2020	CSPDarknet-backed PyTorch Implementation	High Inference Speed & Easy Deployment	Not Realeased from Original Authors
v6	2022	Efficient Backbone	Optimized for Industrial Applications & High FPS on Specific Hardware	Limited Open Source Adoption
v7	2022	Extended Efficient Layer Aggregation Networks (E-ELAN) Real-time Pose Estimation	Improved Accuracy and Speed State-of-the-Art Performance	Higher Computational Demands Increased Complexity
v8	2023	New Backbone and Neck Architecture Anchor-free Detection Advanced Training Strategies Real-time Pose Estimation	Enhanced Performance Simplified Model Structure Better Generalisation Flexability and Modularity	Still Maturing Community Support with Limited Documentation
v9	2023	GELAN and PGI	Improved Gradient Flow and Superior Benchmark Accuracy	Higher Computational Demands

2. Research Design

The precise identification and analysis of combat athletes were performed using advanced CV techniques, with all models trained on a T4 graphics processing unit (GPU) via Google Colab (NVIDIA, 2018). Image preprocessing techniques such as cropping, flipping, rotation, translation, brightness adjustment, contrast enhancement, colour augmentation, and saturation adjustment were used to improve high-resolution raw images. The study's objectives include tracking multiple elements in combat sports, implementing an algorithm for classifying combat sports actions, and designing an algorithm that uses real-time data for precise scoring and comprehensive performance analysis of athletes. Unlike other methods of HAR, this approach does not use pose estimation. YOLOv5, YOLOv8, and YOLOv9 were selected to evaluate various iterations of the same model, aiming to identify challenges and enhance further research. The accuracy of positive predictions was measured using precision, relevant instances were detected with recall, and average precision across all classes was determined using mean Average Precision (mAP). Real-time performance was measured in frames per second (FPS), with computational efficiency taking precedence in this study due to the inherently efficient nature of YOLO models. Extensive training was conducted with the YOLOv5 model, utilizing an image size of 640, a batch size of 48, and 500 epochs. To accelerate the training process, caching was enabled. The medium model for YOLOv5/v8/v9 is selected by default during the experiments. However, it is worth noting that performance gains can be obtained by selecting the larger or extra-large variants of these models, assuming adequate hardware resources are available.

2.1 Dataset Collection and Pre-Processing

Dataset collection involved curating diverse images from publicly available sources and self-recorded footage to encompass various combat sports scenarios, including boxing, MMA, and kickboxing. A combination of YouTube videos and live recordings was used to ensure variability in athlete movements, equipment configurations, and environmental conditions. The videos featured various athlete actions such as punches, kicks, defensive manoeuvres, and equipment such as gloves and headgear. Frames were extracted from these videos at specific intervals to capture critical moments and a broad spectrum of movements. This approach resulted in a dataset of 5,261 images. To maintain consistency and avoid bias, the dataset was divided into 70% for the training set, 20% for the test set, and 10% for the validation set, as shown in Figure 1. Pre-processing was an essential step in preparing the dataset for training. Automated orientation correction algorithms were implemented to ensure image consistency, rectifying any rotational discrepancies inherent in the raw video frames. All images were resized to a uniform dimension of 640×640 pixels to standardise input across all models, facilitating a fair comparison and efficient batch processing. Contrast enhancement techniques were applied to improve image quality, ensuring that features relevant for detection and recognition were pronounced. Various augmentation methods were used to resolve class imbalances in the dataset. Images were randomly cropped at 0%–15% zoom.

Rotations of -3° to +3°, horizontal shear of ±6°, and vertical shear of ±3° were used to replicate various viewing angles and perspectives. For various illumination, 10% of pictures were grayscaled. Adjustments were performed

to hue (±4°), saturation (±5%), brightness (±3%), and exposure (±9%) to simulate different lighting conditions. Sensor noise and motion blur in video frames were simulated by blurring up to 2 pixels and adding noise up to 0.15%. Three additional outputs per training sample increased dataset diversity and model generalisation. Combat sports data performance increased with pre-trained YOLO model weights and transfer learning from large datasets. Training data augmentation reinforced model overfitting resistance. YOLOv5 and v8 used existing repositories, whereas v9 was adjusted for the framework. Grid search optimised learning rate, batch size, and epochs. Early stopping and learning rate schedulers optimised training time and prevented overfitting, guaranteeing models converged without further computation. This dataset reflects combat sports' dynamic and varied nature by merging YouTube and live frames with considerable pre-processing and augmentation.

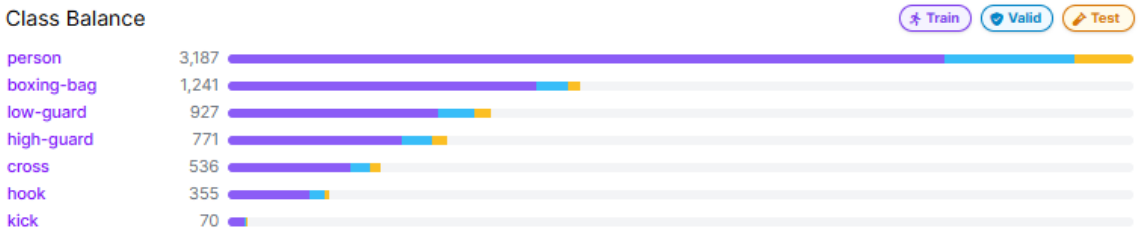


Figure 1: Dataset Classes Balance Breakdown

2.2 Experimentation

The objective is to evaluate the performance of YOLOv5, YOLOv8, and YOLOv9 in detecting objects such as athletes, gloves, and rings by using diverse combat sports datasets that include various scenes and conditions, and then apply standard object detection benchmarks and metrics, including mean Average Precision (mAP), and Precision-Recall metrics. The evaluation compared each model's detection accuracy, ability to handle occlusions, and varying object scales. A multi-object tracking algorithm using detections from each YOLO model was implemented and evaluated. This involved developing a tracking pipeline that integrates these detections within a multi-object tracking framework and evaluating tracking performance using metrics such as accuracy and the number of identity switches. During the evaluation process, particular emphasis was placed on assessing the stability of tracking across both slow-paced and fast-paced sequences, analysing the frequency and causes of identity switches to understand the algorithm's robustness in different conditions. Feature extraction was used to detect objects and extract relevant features for HAR to recognise human actions. The classifier used machine learning (ML) classifiers to analyse the detected features and identify actions. Performance was evaluated using metrics for accuracy, precision, and recall. During the evaluation process, the models were tested on labelled datasets of combat sports actions, analysing performance across different action categories such as punches, kicks, and defensive manoeuvres.

2.3 Computational Efficiency

The computational efficiency of each YOLO model was evaluated by measuring the average inference time per image to determine real-time applicability and memory consumption during model execution. The size of each model was compared to assess storage requirements, and post training tests were performed on various hardware setups, including high-performance GPUs, to understand the impact of the computer hardware on model performance. Computational metrics were recorded and analysed to identify the most efficient model for deployment scenarios. The experimental setup facilitated a comprehensive evaluation of YOLOv5, YOLOv8, and YOLOv9 models across three key areas: object detection, tracking, and HAR. For object detection, the models were assessed on their ability to accurately identify and localise athletes and equipment in different combat sports scenarios using mAP to provide a detailed performance analysis.

2.4 Experimental Setup

All experiments were conducted using Google Colab Pro, which provides access to high-performance GPUs. The hardware and software specifications are shown in Table 2. Additional libraries including NumPy, OpenCV, Matplotlib, and other dependencies as required by each YOLO implementation, were also used.

Table 2: Hardware and Software Specifications

Hardware		Software	Version
GPU	NVIDIA Tesla T4 with 16 GB GDDR6 memory	Ubuntu OS	18.04.6 LTS
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz	Python	3.8.16
RAM	25GB	PyTorch	1.13.1+cu116
		CUDA	11.6
		NVIDIA GPU Drivers	SMI 510.47.03

3. Results

3.1 Model Training Results

Table 3 shows YOLOv5, v8, and v9 training computational performance characteristics. These measurements show model differences in training length, resource use, and final model size. YOLOv8's reasonable model size and low GPU memory suit it for low-GPU situations. YOLOv5, the smallest model, is appropriate for devices with low storage. YOLOv9's memory limits required batch size and epoch modifications, which increased training time and model size.

Table 3: Computational Performance Metrics for YOLO Models

Model	Epochs (Completed)	Training Duration (Hours)	Peak GPU Memory Usage (GB)	Model Size (MB)
YOLOv5	214	1.866	12.7	40.3
YOLOv8	189	2.069	7.28	49.6
YOLOv9	99	4.092	12.9	49

YOLOv5 showed robustness in detecting static objects, while YOLOv8 excelled in scenarios involving rapid movements. YOLOv8's improved stability was particularly evident in sequences with fast-paced actions, improving overall tracking performance. Performance metrics for each model are presented in Tables 4, 5, and 6 respectively, and include accuracy, precision, and recall, providing a comprehensive evaluation of each model's effectiveness in HAR. The computational efficiency of each model was thoroughly tested by measuring inference time, memory usage, and model size. These tests revealed YOLOv5 as the most efficient model for resource-constrained environments, while YOLOv8 balanced performance and computational requirements, making it suitable for real-time applications. The training process included a patience value of 100 for early stopping and halting training early due to no improvements over the last 100 epochs, with the best performance observed at epoch 113. The results summary showed 214 epochs completed over a training duration of 1.866 hours, using 21.5G of memory. Performance metrics for YOLOv5 included an average precision (P) of 0.839 and an average recall (R) of 0.597. Table 4 details the training results for YOLOv5.

Table 4: YOLOv5 Training Results

Class	Images	Instances	P	R	mAP50	mAP50-95
boxing-bag	500	309	0.99	0.83	0.894	0.78
cross	500	105	0.82	0.72	0.772	0.481
high-guard	500	192	0.77	0.75	0.76	0.399
hook	500	92	0.65	0.16	0.201	0.117
kick	500	8	1	0	0	0
low-guard	500	213	0.74	0.77	0.725	0.434
person	500	663	0.92	0.96	0.967	0.763

Similar to YOLOv5, early stopping was used for the YOLOv8 training results, saving the best model at epoch 89 after no improvements were seen in the last 100 epochs, using the model yolov8m.pt (medium) with 250 epochs and an image size of 640. The results summary showed 189 epochs completed during a training duration of 2.069 hours, utilising 7.28G of memory. Performance metrics included precision (P) of 0.704, recall (R) of 0.564,

mAP50 of 0.639, and mAP50-95 of 0.451, reflecting its balanced performance across various evaluation criteria. YOLOv8 training results are detailed in Table 5.

Table 5: YOLOv8 Training Results

Class	Images	Instances	P	R	mAP50	mAP50-95
boxing-bag	500	309	0.99	0.58	0.888	0.821
cross	500	105	0.78	0.69	0.744	0.489
high-guard	500	192	0.74	0.65	0.71	0.374
hook	500	92	0.29	0.14	0.193	0.0914
kick	500	8	0.53	0.25	0.259	0.146
low-guard	500	213	0.71	0.68	0.709	0.454
person	500	663	0.88	0.96	0.972	0.779

The YOLOv8 model demonstrated strong performance in identifying the 'boxing-bag' and 'person' classes but exhibited lower accuracy in detecting the 'hook' and 'kick' classes, mirroring the results for YOLOv5. Training with yolov8_large or other variants may enhance performance and accuracy, although improving the dataset is important to improving the performance of detections on under-represented classes. YOLOv9 was trained with a batch size of 48, 150 epochs, and an image size of 640, using pre-trained weights gelan-c.pt. The YOLOv9 `gelan-c.pt` model leverages the GELAN to enhance object detection performance, offering efficient and accurate inference capabilities suitable for various computational environments. The training process encountered GPU memory constraints necessitating adjustments in batch size and epochs. The results summary indicated 99 epochs completed over a training duration of 4.092 hours, using 12.9G of memory. Performance metrics included average precision (P) of 0.661, average recall (R) of 0.606, average mAP50 of 0.637, and average mAP50-95 of 0.456. Table 6 details the training results for YOLOv9 and shows strong precision for 'boxing-bag' and 'person' classes but deficiencies in 'kick' and 'hook' detection, suggesting potential gains from dataset augmentation or hyperparameter adjustments.

Table 6: YOLOv9 Training Results

Class	Images	Instances	P	R	mAP50	mAP50-95
boxing-bag	500	309	0.99	0.85	0.915	0.843
cross	500	105	0.78	0.71	0.763	0.516
high-guard	500	192	0.77	0.76	0.784	0.429
hook	500	92	0.39	0.21	0.235	0.116
kick	500	8	0	0	0.0257	0.0103
low-guard	500	213	0.77	0.75	0.769	0.461
person	500	663	0.92	0.96	0.968	0.818

3.2 Model Performance

YOLOv5 and YOLOv8 displayed robust performance with high precision in detecting the 'boxing-bag' and 'person' classes, suggesting that these models are well-suited for tasks involving these objects. However, all models struggled with the 'kick' and 'hook' classes, indicating a need for dataset augmentation or further hyperparameter tuning. YOLOv8 showed an efficient training process with balanced GPU memory utilisation and training time. YOLOv9 encountered memory issues requiring batch size and epochs adjustments, leading to longer training times. Early stopping optimised training durations and prevented overfitting, indicating that better detection of 'hook' and 'kick' requires more representative data. Data augmentation and synthetic data generation could enhance model performance for these underrepresented classes. YOLOv9 maintained high precision but had lower recall for specific classes, indicating areas for improvement. While YOLOv5 excelled in precision and recall for key classes, YOLOv8 offered balanced performance, underscoring the need for targeted enhancements in combat sports detection. Implementing early stopping was important in preventing overfitting and saving training time. Adjusting the patience parameter can further optimise training durations without compromising model performance.

The low performance in detecting certain classes, such as 'hook' and 'kick' across all models, highlights the necessity for targeted improvements, either through enhanced training data or refined model architectures. The dataset has the lowest number of 'kick' and 'hook', which suggests that the dataset could be increased to improve performance. Early stopping significantly optimises training efficiency by mitigating overfitting, emphasising the need for more representative data to improve 'kick' and 'hook' class detection. The results from all three models are displayed in Table 7, where the YOLOv5 model achieves high recall and precision for the terms 'boxing-bag' and 'person,' indicating robust performance across multiple classes. In comparison, YOLOv8 offers a balanced trade-off between precision and recall, with notable improvements in detecting 'cross' and 'high-guard' classes. Meanwhile, YOLOv9 maintains high precision but experiences a decline in recall for specific classes such as 'kick' and 'hook', highlighting the need for dataset augmentation or further tuning to enhance its performance in these areas. Overall, the results indicate that while each model has its strengths, there are specific areas where targeted improvements can significantly enhance their effectiveness for combat sports. Given that the dataset is unbalanced and has potential for improvement, the object detection on the validation batch of the three models performed better than anticipated in Figures 2, 3, and 4. YOLOv9 has superior precision and mAP50 and mAP50-95 for 'person' and 'boxing-bag' classes than v8. High-occlusion sequences test YOLOv9's track consistency. YOLOv9 needs more focused data or architecture improvements due to higher memory utilisation, longer training times, and low 'hook' and 'kick' performance.

Table 7: Comparison of YOLOv5/v8/v9 model's metrics

Model	Images	Instances	P	R	mAP50	mAP50-95
YOLOv5	500	1582	0.84	0.6	0.617	0.425
YOLOv8	500	1582	0.7	0.56	0.639	0.451
YOLOv9	500	1582	0.66	0.61	0.637	0.456



Figure 2: YOLOv5 Prediction Results from Validation Batch



Figure 3: YOLOv8 Prediction Results from Validation Batch

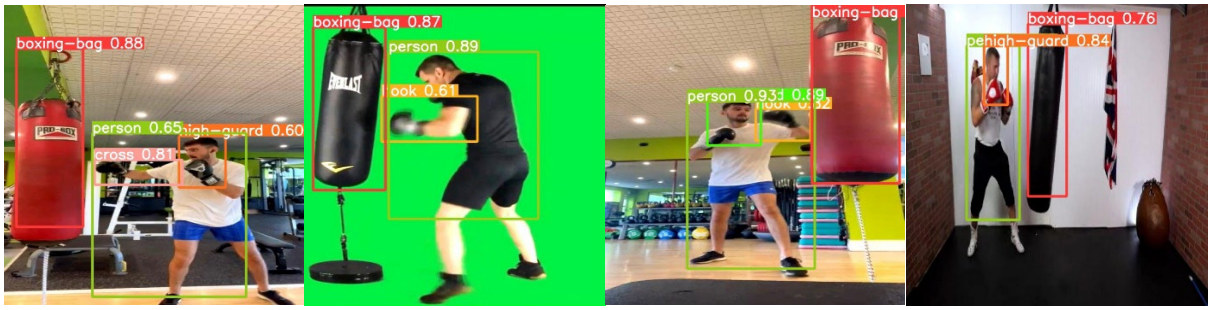


Figure 4: YOLOv9 Prediction Results from Validation Batch

The results provide a detailed breakdown of each model's accuracy in classifying actions, highlighting strengths and areas of misclassification. This pattern of misclassification underscores persistent challenges within dataset and model training processes, particularly for 'hook' and 'kick' actions, necessitating expanded, diverse datasets and enhanced annotation quality for more precise model learning. Addressing these issues is essential for improving model performance, reducing confusion, and enhancing accuracy in recognising complex actions in real-world scenarios. Figure 5 displays scenes from a sample video that contained some frames used to train the model.

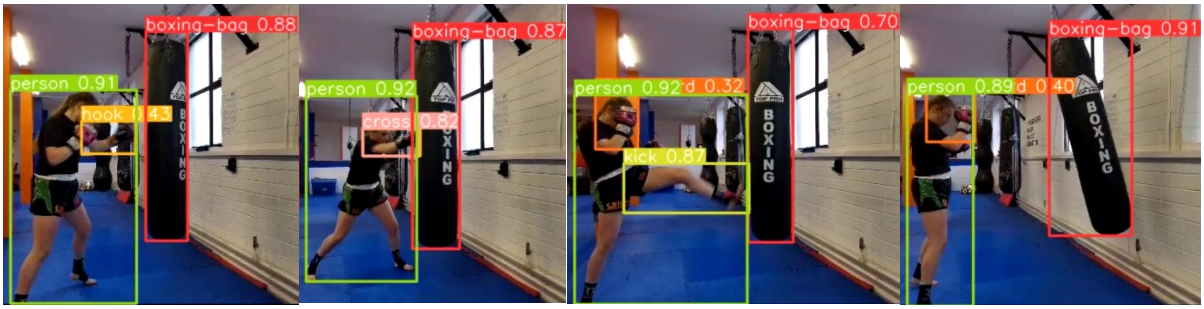


Figure 5: YOLOv8 Results from Video Sample - Hook, Cross, Kick, High Guard, Person, and Boxing-Bag Classes

4. Discussion

Comparing YOLOv5, v8, and v9 in combat sports shows their strengths and weaknesses. Precision and recall are YOLOv5's strengths in 'person' and 'boxing-bag' detection and tracking. Memory utilization and training duration efficiency make it suitable for resource-constrained applications. It misclassifies 'hook' and 'kick' actions, requiring better training data or model optimisation. The balanced YOLOv8 method provides moderate class precision and recall. It identifies 'cross' and 'high-guard,' enhancing tracking stability during rapid movements. Despite these benefits, issues with the 'hook' and 'kick' classes, similar to YOLOv5, suggest dataset limits rather than model architecture problems. YOLOv9 stresses precision but has unpredictable recall, especially for complex categories such as 'hook' and 'kick.' This sophisticated architecture uses GELAN and PGI, resulting in high memory usage and long training times. Improvements in detection accuracy may need additional processing resources, making YOLOv9 more suitable for hardware-capable applications. Dataset class imbalance is seen in all models' inability to detect 'hook' and 'kick' actions. Models cannot learn and generalise these categories without these activities in the training data. This suggests that dataset quality and balance in DL model training is necessary for specific applications. Temporal analysis methods for object detection and tracking may improve combat sports HAR. Transformers and long short-term memory (LSTM) networks improve dynamic action understanding by efficiently managing sequential data and temporal dependencies. These advanced algorithms and YOLO's real-time object identification can improve HAR systems' accuracy and reliability. Coaching and combat sports scoring require quick decisions and feedback, which matches their fast and precise video frame processing.

4.1 Limitations and Future Work

The dataset's class imbalance, especially the under-representation of 'hook' and 'kick' activities, limits this study. The models' poor class detection may have been due to this imbalance. Data augmentation or video of under-represented actions should balance the dataset. Because of its processing needs, YOLOv9 may not be suited for

all applications. This can be solved via model optimisation or hardware acceleration. Future studies could combine temporal analysis with object detection algorithms to improve HAR, possibly using time series analysis.

5. Conclusions

This study examines combat sports object detection, tracking, and HAR using the YOLOv5, YOLOv8, and YOLOv9 object detection systems. A comparative analysis of the three models selected YOLOv8 due to its balanced precision and recall across classes, consistent object identification, and robust performance measures. Its capacity to improve tracking stability during quick movements makes it ideal for coaching, training, and competitive scoring. YOLOv5 accurately detects 'person' and 'boxing-bag' classes but not 'hook' or 'kick' actions. Its training duration and memory economy make it suitable for resource-constrained applications. YOLOv9's precision and superior architecture may increase detection rates. However, it demands more processing resources and sometimes fails to keep consistent tracks in high-occlusion or overlapping action sequences, limiting its applicability. The results show that YOLO models effectively detect and track combat sports object identification, emphasising the need for model selection based on application needs. The difficulty of models detecting specific action classes suggests dataset restrictions must be addressed.

Additional data must be gathered and enhanced to represent under-represented categories, such as ‘hook’ and ‘kick' to improve detection accuracy and efficiency. Enhancing hyperparameters and model architectures should elevate performance in difficult classes without augmenting computational demands. Integrating temporal analysis for HAR is enhanced when YOLO models are combined with temporal networks to capture the temporal dependencies of action sequences. These advancements can enhance HAR systems and the identification of objects in combat sports, thereby improving sports analytics and technology-driven coaching. This study aims to enhance the understanding of CV and AI technologies for sports analytics by examining cutting-edge object detection models within the specific domain of combat sports. The findings can be used to improve training methodologies, monitor athletes' movements and actions, tailor coaching interventions, and increase the objectivity and consistency of competitive scoring.

References

Angelini, F., Fu, Z., Long, Y., Shao, L. and Naqvi, S. M. (2019) '2D Pose-Based Real-Time Human Action Recognition with Occlusion-Handling', *IEEE Transactions on Multimedia,* 22(6), pp. 1433-1446.

Atik, M. E., Duran, Z. and Özgünlük, R. (2022) 'Comparison of YOLO versions for Object Detection from Aerial Images', *International Journal of Environment and Geoinformatics,* 9(2), pp. 87-93.

Binter, J., Říha, D., Boschetti, S., Prossinger-Beck, V., Stachoň, M. and Prossinger, H. (2024) 'Navigating Between Human and Machine-Based Evaluation: Judgment and Objectivity in Economic Games Exemplified in the Analysis of MMA Fights': Springer Nature Switzerland, pp. 235-246.

Bochkovskiy, A., Wang, C. Y. and Liao, H. Y. M. (2020) 'Yolov4: Optimal Speed and Accuracy of Object Detection', *arXiv preprint,* 1(arXiv:2004.10934), pp. 1-17.

Boesch, G. (2023) *YOLOv7: A Powerful Object Detection Algorithm (2024 Guide)*: viso.ai. Available at: https://viso.ai/deep-learning/yolov7-guide/ 2024).

Echeverria, J. and Santos, O. C. 'Punch Anticipation in a Karate Combat with Computer Vision', *UMAP '21: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, Utrecht, Netherlands, 21-25 June: ACM, pp. 61–67.

Host, K. and Ivašić-Kos, M. (2022) 'An Overview of Human Action Recognition in Sports Based on Computer Vision', *Heliyon,* 8(6), pp. 1-25.

Ishac, K. and Eager, D. (2021) 'Evaluating Martial Arts Punching Kinematics Using a Vision and Inertial Sensing System', *Sensors,* 21(6), pp. 1-25.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D. and Makedon, F. (2020) 'A Survey on Contrastive Self-Supervised Learning', *Technologies,* 9(1), pp. 1-22.

James, L. P., Robertson, S., Haff, G. G., Beckman, E. M. and Kelly, V. G. (2017) 'Identifying the Performance Characteristics of a Winning Outcome in Elite Mixed Martial Arts Competition', *Journal of Science and Medicine in Sport,* 20(3), pp. 296-301.

Jiang, P., Ergu, D., Liu, F., Cai, Y. and Ma, B. (2022) 'A Review of Yolo Algorithm Developments', *Procedia computer science,* 199(1), pp. 1066-1073.

Kumar, A. (2022) 'Gamification in Training with Next Generation AI-Virtual Reality, Animation Design and Immersive Technology', *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1-14.

Liau, Y. Y. and Ryu, K. (2021) 'Status Recognition Using Pre-Trained YOLOv5 for Sustainable Human-Robot Collaboration (HRC) System in Mold Assembly', *Sustainability,* 13(21), pp. 1-15.

Makovsky, Z. (2020) 'The 10-point system - a must? A critique of MMA’s scoring system', *The Fight Site*. Available at: https://www.thefight-site.com/home/10-point-system-critique-scoring-mma 2024].

Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C. and He, Z. 'Spatially supervised recurrent convolutional neural networks for visual object tracking'. *2017 IEEE international symposium on circuits and systems (ISCAS)*, Baltimore, MD, USA, 28-31 May: IEEE, pp. 1-4.

NVIDIA (2018) *NVIDIA Tesla T4 GPU Architecture*: NVIDIA. Available at: https://www.nvidia.com/en-us/data-center/tesla-t4/.

Pang, Y., Wang, Q., Zhang, C., Wang, M. and Wang, Y. 'Analysis of Computer Vision Applied in Martial Arts', *2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China, 14-16 January. https://ieeexplore.ieee.org/xpl/conhome/9712655/proceeding: IEEE, pp. 1-6.

Patil, A. and Rane, M. 'Convolutional Neural Networks: An Overview and its Applications in Pattern Recognition', *Information and Communication Technology for Intelligent Systems*, Online, May 15-16. Singapore: Springer Nature, pp. 21-30.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. 'You Only Look Once: Unified, Real-Time Object Detection', *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 1 January: CVPR, pp. 779-788.

Saponara, S. (2017) 'Wearable Biometric Performance Measurement System for Combat Sports', *IEEE Transactions on Instrumentation and Measurement,* 66(10), pp. 1-11.

Sohan, M., Sai Ram, T., Reddy, R. and Venkata, C. 'A review on yolov8 and its advancements'. *International Conference on Data Intelligence and Cognitive Informatics*: Springer, pp. 529-545.

Terven, J. and Cordova-Esparza, D. (2023) 'A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond', *arXiv preprint arXiv:2304.00501,* 1(1), pp. 1-34.

Van Zandycke, G., Somers, V., Istasse, M., Del Don, C. and Zambrano, D. (2022) 'DeepSportradar-v1: Computer Vision Dataset for Sports Understanding with High Quality Annotations', *MMSports '22: 5th International ACM Workshop on Multimedia Content Analysis in Sports,* 5(5), pp. 1-8.

Veit, A., Matera, T., Neumann, L., Matas, J. and Belongie, S. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140.* IEEE Xplore.

Wang, C.-Y., Yeh, I.-H. and Liao, H.-Y. M. (2024) 'Yolov9: Learning What you Want to Learn Using Programmable Gradient Information', *arXiv preprint arXiv:2402.13616*.

Yang, G., Feng, W., Jin, J., Lei, Q., Li, X., Gui, G. and Wang, W. 'Face mask recognition system with YOLOV5 based on image recognition'. *IEEE 6th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 11-14 December: IEEE, pp. 1398-1404.

Yap, M. H., Hachiuma, R., Alavi, A., Brüngel, R., Cassidy, B., Goyal, M., Zhu, H., Rückert, J., Olshansky, M. and Huang, X. (2021) 'Deep learning in diabetic foot ulcers detection: a comprehensive evaluation', *Computers in Biology and Medicine,* 135(104596), pp. 1-17.

Yin, Y., Li, H. and Fu, W. (2020) 'Faster-YOLO: An accurate and faster object detection method', *Digital Signal Processing,* 102(1), pp. 1-11.

Yu, S., Wu, G., Gu, C. and Fathy, M. E. 'TDT: Teaching Detectors to Track without Fully Annotated Videos', *CVPR 2022 Workshop: Workshop on Learning with Limited Labelled Data for Image and Video Understanding (L3D-IVU)*, New Orleans, Louisiana, USA, 21 - 24 June: IEEE, pp. 3940-3950.

AI Generative in Brazil's Public Schools: The Teachers' Perspective

Jussara Reis-Andersson
Mid Sweden University, Department of Education, Sundsvall, Sweden

Jussara.reis-andersson@miun.se

Abstract: Artificial Intelligence (AI) alters the societal landscape and transforms people's lives. In the realm of education, AI offers many opportunities and challenges in teaching and learning. While AI holds promise in enriching learning experiences and fostering motivation, its successful integration into educational practice necessitates that both students and teachers possess digital competence. Education plays a critical role in preparing future workforces. In South America, Brazil stands as the most populous country, with 212.6 million people. The utilisation of AI generative tools, such as ChatGPT, for teaching purposes in the country, is on the rise, albeit constrained within the educational context. Challenges faced in integrating AI generative tools in Brazil's education system include the facilitation of plagiarism and teachers' insufficient digital competence. In an effort to delve into the intricacies of teachers' experiences and challenges regarding the integration of AI generative tools into their pedagogical practice, a qualitative study was conducted, and nine teachers in an upper secondary school in Brazil answered a questionnaire involving seven open questions. The teachers perceive the integration of AI generative tools into teaching practice as an instrument that enhances opportunities for improving the quality of teaching, stimulating student interest, and enriching the dynamism of the content learning process. Teachers have highlighted ChatGPT as a valuable tool for research and consultation. This tool can simplify teaching tasks and is typically well-received by students. Both teachers and students can utilise this tool to generate materials such as lyrics and textual variations and even produce fully composed songs ready for listening. ChatGPT is also used to prepare and correct assessments, ensuring consistency in evaluation. Teachers use AI to prepare texts, quizzes, and multiple-choice tests. AI-generated images are also used to enhance illustrations, making learning materials more engaging and visually appealing. Although AI simplifies complex subjects and makes the learning process more engaging, it is essential to provide adequate training on AI tools to enhance student involvement and educational results.

Keywords: Artificial intelligence, Developing country, Education, School, Teacher, Teaching

1. Introduction

In recent years, the use of artificial intelligence (AI) has increased worldwide, creating opportunities and challenges in many areas, such as education. This is an important standpoint because, as Freire (2020) explained, education helps people find their voices and become more conscious, enabling them to reach their full potential by speaking, thinking, and acting freely. In developing countries, access to and use of AI generative technologies may be limited due to a lack of financial opportunities, well-functioning infrastructure, and teachers' skills in using AI as an educational tool. Yet, AI generative technologies are a reality that cannot be ignored.

In the realm of education, AI brings a host of opportunities and challenges to teaching and learning. Although AI can significantly enhance learning experiences and increase motivation, successful integration into educational practices depends on both students and teachers having strong digital skills (Kuka et al., 2022). However, de Oliveira Durso and Arruda (2022) explain that perceiving AI as having independent intentions is unrealistic.

AI is becoming more advanced, changing how education works, which increases the need for understanding how this tool can be used in teaching and learning practices (Gunnars, 2024). This paper intends to understand how teachers in Brazil's upper secondary schools perceive AI in teaching practice and the opportunities and challenges related to the integration of AI in teaching. The following research questions are put forward:

RQ1: How do teachers in Brazil's upper secondary schools perceive the integration of AI generative tools in teaching?

RQ2: What opportunities and challenges related to the integration of AI generative tools do teachers in Brazil's upper secondary schools describe?

2. Background

Artificial intelligence (AI) is transforming the world, people's lives, and the education systems. This technology provides numerous opportunities, but it also brings challenges that we must address. For instance, it is not just about using AI; it is about how we use this tool effectively for teaching and learning. Silva (2022) highlights a shift in the pedagogical process where techniques play a decisive role in shaping educational outcomes.

Education is crucial in preparing workforces for an uncertain future. The integration of AI tools not only involves adopting advanced technologies to enhance learning and address the AI skills gap but also requires a reformation in educational approaches (Pedro et al., 2019). From the perspective of developing countries, understanding

teachers' views on using AI in their teaching practice is crucial. These teachers face unique challenges when integrating this technology into their classrooms. Brazil, the most populous country in South America with about 215 million people, serves as an important example. According to Carius (2021), the use of digital technologies in schools in Brazil remains precarious. The Brazilian education system faces challenges in integrating these AI tools include facilitating plagiarism, teachers' lack of digital skills, and worsening educational inequality (Godoi, 2024). Exploring how teachers in Brazil integrate AI tools into their teaching practice and the challenges they encounter can offer valuable insights. By understanding these opportunities and challenges, we can better support teachers worldwide in harnessing AI to enhance education for all. In light of this context, AI rises as a transformative force with the potential to revolutionise educational practices, making learning more engaging and accessible. However, realising this potential requires addressing the associated challenges and ensuring that both teachers and students are adequately prepared to utilise AI effectively.

Russell and Norvig (2016) describe AI as an intelligent system capable of maximising task success rates. This technology presents significant opportunities for enhancing teaching and learning practices, such as personalised learning experiences and more efficient administrative tasks. However, AI distinguishes itself from other classroom technologies by its autonomous and automated approach to performing tasks (DURSO, 2024). Lindqvist and Arvidsson (2024) emphasise that teachers must critically assess AI resources and tailor them to fit their educational contexts.

To address challenges, the use of AI generative technology in education needs structured training, support, and clear policy guidelines for both students and teachers (Albadarin et al., 2024; Adeshola and Adepoju, 2023). According to Albadarin et al. (2024), more research and proactive dialogues with policymakers, stakeholders, and educational practitioners are needed to refine and enhance AI use in teaching and learning practices.

2.1 ChatGPT in Education

The term ChatGPT stands for Chat Generative Pre-training Transformer, an AI chatbot based on large language models (LLM) (Zhao et al., 2023). This online text-based AI chatbot can interact with humans and generate human-like responses. In a literature review, Mai et al. (2024) identified 51 articles on the use of ChatGPT in educational practices. Only three of these articles were from South America: two from Peru and one from Uruguay.

ChatGPT is one of the several AI tools that have been used in education. Its impact on pedagogical practices and student engagement has been discussed due to the opportunities and challenges that the use of ChatGPT entails. This tool has the potential to revolutionise educational practices, but the challenges connected to its integration need to be taken into account (Mohebi, 2024). Islam and Islam (2024) explored the opportunities and challenges of ChatGPT in education and explained that ChatGPT could offer personalised learning experiences for students but may also provide false information due to mistakes in its training data. Leite (2023) pointed out that ChatGPT can, among other possibilities, write and summarise texts, answer questions, and draw up lesson plans.

There is potential in using ChatGPT in educational practices, but careful integration and adaptation of this tool, along with improved professional development for teachers in utilising ChatGPT and other AI tools in teaching and learning practices, are essential (Al-khresheh, 2024). Teachers recognise ChatGPT's usefulness in various applications, such as lesson preparation and creating learning activities, but they also express concerns that students may become dependent on ChatGPT and question their own knowledge (Ulla et al., 2023). Using ChatGPT in teaching practice may enhance students' motivation and satisfaction with learning processes (Woo et al., 2024). However, according to Al-khresheh (2024), teachers should allow students enough time to enhance their engagement with materials and tasks. Students need to use critical thinking when using ChatGPT in learning processes, and teachers can use ChatGPT to generate responses as models for developing students' critical thinking skills (Janse van Rensburg, 2024). Jayaron and Blessy Jayaron (2024) argue that further research is necessary to understand the implications of tools like ChatGPT and to develop effective anti-cheating strategies.

Sandu et al. (2024) pointed out that the discourse surrounding ChatGPT in education settings encompasses a range of perspectives, highlighting the pedagogical utility, ethical considerations, and potential challenges. The scholars emphasised that while ChatGPT provides personalised and flexible support for students and teachers, it is limited in understanding complex queries and lacks human interaction. However, frequent users of ChatGPT perceive more benefits from this tool than those who do not use it (Sandu et al., 2024).

3. Method

This study adopted a qualitative research design to facilitate the extraction of significant insights from the data (Cohen et al., 2017). The data were collected by distributing a questionnaire to thirty teachers in an upper secondary school in Brazil. Only nine teachers, which is about 30 percent of the respondents, have answered the questionnaire. This low response rate in the questionnaire can be addressed by elements such as teachers' workloads and the time the questionnaire was distributed, May and June, before the mid-year holidays.

3.1 The Questionnaire Construction

The questionnaire comprised a total of fourteen questions, divided into two sections: seven background questions and seven open questions. The background questions gather information about the respondents, such as their teaching experience, subject areas, and familiarity with AI technologies. Nine teachers, four men, and five women answered the questionnaire. Over 50 % were aged 30-50 and had 5-15 years of experience. All nine teachers had received teacher training and were qualified to teach their subjects.

The open questions explored the perceptions, opportunities, and challenges associated with the integration of AI, specifically ChatGPT, into educational practices. These questions aimed to understand how teachers perceive the impact of AI on their teaching practices, the ways they are currently using AI tools in their classrooms, and the opportunities they have observed from such integration.

3.2 Data Analysis

Inspired by Braun & Clark, the data was unpacked and analysed thematically. First, I familiarised myself with the data by reviewing the collected data. After that, I organised the data to generate codes and search for themes. For example, AI, such as ChatGPT, helps prepare questions and update the creation of assessments. These two codes are connected because they are about AI as a transformative tool in teaching practice. It is important to note that, these codes are not strictly limited to a single theme. After the codes were generated and themes revised, I clustered codes into themes. Five themes were found: AI as a transformative tool, AI in assignments and tests, efficiency and engagement, challenges in AI integrations, and evaluating student engagement. To strengthen the data, some quotations were included.

3.3 Method Discussion

Due to the short timeframe and difficulty scheduling teachers, I decided to send the questions by e-mail. This approach allowed teachers to respond at their convenience, as they could choose the most suitable time to answer the questions. However, a drawback is that I could not observe the teachers' facial expressions and body language, which are visible during face-to-face interviews. This lack of visual cues limits interpreting non-verbal communication.

4. Findings

This section presents the findings of this study. Five themes have been founded: AI as a transformative tool, AI in assignments and tests, efficiency and engagement, challenges in AI integrations, and evaluating student engagement. The findings may not be universally generalisable, but they improve the knowledge of the importance of perceptions of AI tools in teaching practice. Table 1 summarises the findings of this paper.

Table 1: Summary of AI impact on education: key themes, descriptions and teacher quotations

Theme	Description	Teacher Quotation
AI as a transformative tool	AI improves teaching, creates assessments, and boosts participation.	"AI can be used in my didactics and in my working methodologies." "AI can be used for research and consultation, which is excellent. It can produce materials such as lyrics, variations of texts, and even the final version of a song ready to be listened to."
AI in assignments and tests	AI aids in preparation, grading ensuring consistency in evaluations and activities.	"AI can be used for correcting assessments, formulating activities and, above all, checking the activities carried out by the students." "AI can be used to prepare texts, quizzes, and tests with alternatives A, B, C and D."

Theme	Description	Teacher Quotation
Efficiency and engagement	AI increases efficiency, boosts participation, and optimises teaching time.	"AI offers agility and speed in processing information." "AI is a tool for time optimisation at work."
Challenges in AI integration	Meeting individual needs, verifying sources, and adapting to new tech.	"A challenge is to recognise each student's difficulties and check the legitimacy of the sources." "A challenge is teachers' mastery of technology."
Evaluating student engagement	AI simplifies learning but risks superficial understanding; training is essential.	"It is easier to teach." "We can see that young people are very connected to technology. There is a lot of interest in using computers, perhaps not yet in terms of studies, but in general contexts such as games, social networks, and various websites."

4.1 AI as a Transformative Tool

Teachers perceive generative AI as a transformative tool in their teaching practice, offering several opportunities and improvements. They noted that AI helps in preparing questions, updating the creation of assessments, and adapting them to different learning levels. One teacher pointed out, "AI can be used in my didactics and in my working methodologies." Specific areas where AI has been helpful include enhancing both student learning and teaching practice, particularly in active participation and engagement strategies. Another teacher highlighted AI's role in research and consultation, noting how it assists in producing materials like song lyrics, text variations, and complete songs, thereby enhancing the creative aspects of teaching and providing diverse learning resources. The teacher explained "AI can be used for research and consultation, which is excellent. It can produce materials such as lyrics, variations of texts, and even the final version of a song ready to be listened to." This capability highlights how AI tools, including ChatGPT, Google Search, and YouTube enhance the potential for improving learning outcomes. By making information more accessible, these tools trigger student interest and add dynamism to the classroom experience.

4.2 AI in Assignments and Tests

Teachers frequently utilise AI tools for assignments and tests, highlighting their flexibility in educational practices. These tools play a crucial role in both preparing and grading assessments, ensuring a consistent evaluation process across students. One teacher emphasised that "AI can be used for correcting assessments, formulating activities and, above all, checking the activities carried out by the students." This use of AI not only saves teachers time but also facilitates more frequent and varied assessments, thereby enhancing the overall efficiency and effectiveness of the evaluation process.

In the realm of content creation, AI proves invaluable for generating texts, quizzes, and multiple-choice tests. These AI-generated materials are adapted to meet teaching objectives while maintaining relevance. Moreover, according to teachers, AI extends its utility to enhancing visual learning benefits through the creation of images and illustrations. Another teacher explained, "AI can be used to prepare texts, quizzes, and tests with alternatives A, B, C and D." These visuals not only reinforce learning concepts but also capture students' attention, thereby enhancing their overall engagement with the learning materials.

4.3 Efficiency and Engagement

Teachers highlighted the ease and speed provided by AI tools, noting a significant increase in efficiency. AI allows them to accomplish tasks more quickly than traditional methods. One teacher clarified, "AI offers agility and speed in processing information." AI also serves as an extra tool to encourage student integration by offering diverse and innovative ways to engage students, creating a more inclusive and interactive classroom atmosphere. Teachers pointed out that this integration encourages students to participate more actively in their learning process. A further advantage is the optimisation of time. One teacher pointed out, "AI is a tool for time optimisation at work." AI tools help simplify various tasks, making it possible for teachers to focus on essential aspects of teaching, such as personalised instruction and student support. AI also summarises content and provides quick access to information, offering teachers and students many possibilities for deeper exploration and understanding of subjects.

4.4 Challenges in AI Integration

Concerning the integration of AI in teaching, teachers have highlighted several challenges. A significant challenge is addressing each student's difficulties and verifying the legitimacy of sources. One teacher explained, "A

challenge is to recognise each student's difficulties and check the legitimacy of the sources." Teachers also emphasised that traditional teaching methods allow them to interact directly with students, observe their struggles, and provide personalised support. With AI, there is a risk of missing these nuances, as the technology may not always respond precisely to each student's unique needs. Ensuring the information provided by AI is accurate and from credible sources is crucial to prevent misinformation.

Another challenge is adapting to new tools. Many teachers have spent years enhancing their skills with traditional methods and may find it challenging to learn and integrate new technologies. This adaptation process requires time, training, and dedication, which can be problematic for some. There is also the challenge of avoiding too much reliance on AI for lesson preparation. While AI can be a powerful assistant, it is essential for teachers to maintain their creativity and ingenuity. Over-reliance on AI might lead to a decline in these skills and reduce the personalised touch they bring to their classrooms.

Lastly, one teacher pointed out, "A challenge is teachers' mastery of technology." Not all teachers are equally comfortable with or proficient in using new technologies. Bridging this gap is necessary to ensure the effective use of AI tools and that all students benefit from their potential.

4.5 Evaluating Student Engagement with AI

Teachers have shared insights on student engagement with AI tools like ChatGPT in teaching. Some teachers have noted that learning becomes easier with the integration of AI, as these tools can simplify complex topics and provide instant access to information, making it easier for students to grasp new concepts and for teachers to present material engagingly. However, there is also a concern that AI can hinder genuine learning, as students might copy information from AI outputs without understanding the content. This practice undermines the learning process and does not encourage critical thinking or deep understanding.

On the positive side, some teachers have observed that students react well to the use of technology in the classroom. One teacher pointed out, "It is easier to teach." The familiarity and interest students have in technology help to make lessons more engaging and interactive. Yet, students need proper training on effectively using AI tools. Without understanding how these tools work and their appropriate uses, students may continue to use them in ways that do not enhance their learning process. Proper training promotes AI in ways that promote deeper engagement and better learning outcomes.

Interestingly, one teacher mentioned a lack of experience in observing students using AI tools like ChatGPT. The teacher noted, "We can see that young people are very connected to technology. There is a lot of interest in using computers, perhaps not yet in terms of studies, but in general contexts such as games, social networks, and various websites." While young people are connected to technology through games, social networks, and websites, this interest does not always translate directly into educational engagement.

In summary, AI serves as a transformative tool in education, enhancing teaching practice by assisting with question preparation, assessment updates, and the creation of diverse teaching materials, such as song lyrics and text variations. Teachers utilise AI for efficiently preparing and correcting assignments and tests, ensuring consistent evaluation. AI also facilitates the creation of engaging texts, quizzes, and AI-generated images, optimising teaching efficiency and promoting deeper understanding through quick content summarisation. Despite its advantages, integrating AI into education poses challenges, including addressing individual student needs, verifying sources, and adapting to new technologies, which require time and training. Additionally, there is concern about potential over-reliance on AI, which may diminish teachers' creativity and students' critical thinking.

5. Discussion

This paper intends to understand how teachers in Brazil's upper secondary schools perceive AI in teaching practice and the opportunities and challenges related to the integration of AI in teaching. In returning to the research questions, the first question yielded information about how teachers in Brazil's upper secondary schools perceive the integration of AI generative in teaching.

Overall, teachers perceive generative AI as a valuable addition to their teaching tools, which aligns with the findings of Sandu et al. (2024). They are integrating AI into educational practices in innovative ways, such as creating and grading assignments, developing engaging learning materials, and providing immediate support and clarification to students. These applications enhance the efficiency, effectiveness, and interactivity of teaching and learning processes. Teachers see both opportunities and challenges in using AI, like ChatGPT, in

education. While these tools can make teaching easier and students often respond well to the integration of technology, there are concerns about superficial learning practices and the need for proper training. Teachers raised the issue of whether students should use tools like ChatGPT in their learning. At the same time, they emphasised that it is more a question of how and when an AI generative tool should be used, rather than whether they should be used at all. This perspective aligns with Lindqvist and Arvidsson (2024) emphasis on assessing AI resources in specific educational contexts. Teachers also pointed out that more experience is necessary to fully understand and evaluate student engagement with AI in educational settings.

The second research question revealed what teachers in Brazil's upper secondary schools perceive as opportunities and challenges of AI. Teachers identified several advantages of using AI, such as ChatGPT, over traditional teaching methods. These advantages include increased ease and speed, enhanced student engagement, time optimisation, and improved information processing, all contributing to a more effective and dynamic educational experience. These findings align with Leite's (2023) statement that ChatGPT offers many opportunities in education. The integration of AI in assignments and tests also underscores its transformative impact on educational practices, offering teachers innovative tools to optimise learning experiences and promote educational success among students.

However, AI also presents challenges, such as addressing individual student needs, verifying source legitimacy, adapting to new tools, avoiding over-dependence on technology, and ensuring teachers are technologically competent. This relates to Jayaron and Blessy Jayaron's (2024) argument about the need to understand the implications of AI in education and to develop effective anti-cheating strategies. Teachers have emphasised that addressing these challenges is essential for successfully integrating AI into teaching, aligning with Albadarin et al. (2024). While AI simplifies complex topics and makes learning more engaging, proper training in AI tool usage is crucial for enhancing student engagement and learning outcomes.

In summary, while AI makes complex topics easier to understand and learning more engaging, it is crucial to have proper training in using AI tools to improve student engagement and educational outcomes. Teachers recognise both the potential opportunities and the challenges of incorporating AI, like ChatGPT, into their teaching practice. These tools can simplify teaching tasks and are generally well-received by students. However, there are concerns about superficial learning and the necessity for comprehensive training. To fully grasp and assess the effects of AI on education, more structured studies and practical experience are needed to close the gap between students' familiarity with technology and their educational engagement.

5.1 Implications

The practical implications highlight the importance of digital competence among teachers in using AI generative tools. Teachers recognise both the potential opportunities and challenges associated with integrating AI, such as ChatGPT, into their practices. AI generative tools can enhance teaching efficiency, student engagement, and learning outcomes. However, ongoing professional development is essential to equip teachers with the skills and knowledge necessary to fully use AI's potential in enhancing educational experiences. Moreover, fostering digital competence among teachers not only aids in AI tools but also enables them to tailor learning experiences and support diverse student needs.

6. Conclusion

Teachers in Brazil's upper secondary schools perceive generative AI, like ChatGPT, as a valuable addition to their teaching toolkit, offering numerous benefits and improvements. They integrate AI into educational practices innovatively, using it to create and grade assignments, develop engaging learning materials, and provide support to students. These applications significantly enhance the efficiency, effectiveness, and interactivity of the teaching and learning process.

Teachers identify several opportunities and challenges associated with AI integration. AI tools like ChatGPT simplify teaching tasks, improve learning experiences, and promote educational success by enhancing speed, ease, and efficiency in assignments and tests. These tools facilitate rapid information processing and provide interactive learning that engages students. Despite the opportunities, integrating AI presents challenges such as addressing individual student needs, verifying source legitimacy, adapting to new technologies, and avoiding excessive reliance on AI. Balancing technological dependency is crucial to maintaining students' critical thinking. Teachers emphasise the importance of training to use AI effectively and creatively in teaching while ensuring it complements traditional teaching approaches rather than replacing them.

AI simplifies complex topics and makes learning more engaging by quickly processing information and adapting content to different learning levels. However, the effective integration of AI into education requires digital competence and training in AI tool usage to maximise student engagement and improve learning outcomes.

Bridging the gap between students' technological familiarity and educational engagement is crucial and necessitates structured learning and practical experience.

6.1 Future Research

Future research is needed to understand the impact of AI generative tools on teachers' and students' utilisation of this technology in both teaching and learning practices. This area deserves attention due to the potential influence of AI use on teachers' creativity in teaching practice and students' development of critical thinking. Further exploration could shed light on how these tools shape teachers' daily practices and students' learning behaviours, educational performance, and overall educational experiences.

References

al-Khresheh, M. H. 2024. Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence,* 6**,** 100218.

Albadarin, Y., Saqr, M., Pope, N. & Tukiainen, M. 2024. A systematic literature review of empirical research on ChatGPT in education. *Discover Education,* 3**,** 60.

Carius, A. C. 2021. Pós-pandemia de COVID-19, ensino híbrido e inteligência artificial: É a virtualização da escola? *Research, Society and Development,* 10**,** e55610716834.

Cohen, L., Manion, L. & Morrison, K. 2017. *Research methods in education,* London, Routledge, London.

DE OLIVEIRA DURSO, S. & ARRUDA, E. P. 2022. Artificial intelligence in distance education: A systematic literature review of Brazilian studies. *Problems of Education in the 21st Century,* 80**,** 679–692.

Durso, S. D. O. 2024. Reflexões sobre a aplicação da inteligência artificial na educação e seus impactos para a atuação docente. SciELO Brasil.

Freire, P. 2020. Pedagogy of the oppressed. *Toward a sociology of education.* Routledge.

Godoi, J. 2024. REFLEXÕES SOBRE O RELATÓRIO GLOBAL DE MONITORAMENTO DA EDUCAÇÃO E O USO DE CELULARES NAS ESCOLAS [Reflections on the Global Education Monitoring Report and the use of Mobile Phones in Schools]. *Revista SL Educacional,* 6**,** 82.

Gunnars, F. 2024. *Smartbands, artificial intelligence, and special educational needs: Multifarious behaviorist design for primary education.* Mid Sweden University.

Islam, I. & Islam, M. N. 2024. Exploring the opportunities and challenges of ChatGPT in academia. *Discover Education,* 3**,** 31.

Janse Van Rensburg, J. 2024. Artificial human thinking: ChatGPT's capacity to be a model for critical thinking when prompted with problem-based writing activities. *Discover Education,* 3**,** 42.

Jayaron, J. & Blessy Jayaron, J. 2024. Educators' Academic Insights on Artificial Intelligence: Challenges and Opportunities. *Electronic Journal of e-Learning,* 22**,** 59–77.

Kuka, L., Hörmann, C. & Sabitzer, B. 2022. Teaching and Learning with AI in Higher Education: A Scoping Review. *Learning with Technologies and Technologies in Learning: Experience, Trends and Challenges in Higher Education***,** 551–571.

Leite, B. S. 2023. Inteligência artificial e ensino de química: uma análise propedêutica do chatgpt na definição de conceitos químicos. *Química Nova,* 46**,** 915–923.

Lindqvist, M. H. & Arvidsson, C. 2024. Exploring Student and AI Generated Texts: Reflections on Reflection Texts. *Electronic Journal of e-Learning,* 22**,** 52–59.

Mai, D. T. T., Da, C. V. & Hanh, N. V. The use of ChatGPT in teaching and learning: a systematic review through SWOT analysis approach. Frontiers in Education, 2024. Frontiers Media SA, 1328769.

Mohebi, L. 2024. Empowering learners with ChatGPT: insights from a systematic literature exploration. *Discover Education,* 3**,** 36.

Pedro, F., Subosa, M., Rivas, A. & Valverde, P. 2019. Artificial intelligence in education: Challenges and opportunities for sustainable development.

Russell, S. J. & Norvig, P. 2016. *Artificial intelligence: a modern approach*, Pearson.

Sandu, R., Gide, E. & Elkhodr, M. 2024. The role and impact of ChatGPT in educational practices: insights from an Australian higher education case study. *Discover Education,* 3**,** 71.

Silva, P. A. P. 2022. *EdTech e a plataformização da educação [EdTech and the platformisation of education].* Tese de Doutorado em Educação, Universidade do Estado do Rio de Janeiro. http://www.bdtd.uerj.br/handle/1/19281.

Ulla, M. B., Perales, W. F. & Busbus, S. O. 2023. 'To generate or stop generating response': Exploring EFL teachers' perspectives on ChatGPT in English language teaching in Thailand. *Learning: Research and Practice,* 9**,** 168–182.

Woo, D. J., Wang, D., Guo, K. & Susanto, H. 2024. Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies***,** 1–28.

Zhao, w. X., zhou, k., li, j., tang, t., wang, x., hou, y., min, y., zhang, b., zhang, j. & dong, z. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

LLM Supply Chain Provenance: A Blockchain-Based Approach

Shridhar Singh and Luke Vorster
University of KwaZulu-Natal, Westville, South Africa

217008024@stu.ukzn.ac.za
VorsterL@ukzn.ac.za

Abstract: The burgeoning size and complexity of Large Language Models (LLMs) introduce significant challenges in ensuring data integrity. The proliferation of "deep fakes" and manipulated information raises concerns about the vulnerability of LLMs to misinformation. Traditional LLM architectures often lack robust mechanisms for tracking the origin and history of training data. This opaqueness can leave LLMs susceptible to manipulation by malicious actors who inject biased or inaccurate data. This research proposes a novel approach integrating Blockchain Technology (BCT) within the LLM data supply chain. With its core principle of a distributed and immutable ledger, BCT offers a compelling solution to address this challenge. By storing the LLM's data supply chain on a blockchain, we establish a verifiable record of data provenance. This allows for tracing the origin of each data point used to train the LLM, fostering greater transparency and trust in the model's outputs. This decentralised approach minimises the risk of single points of failure and manipulation. Additionally, the immutability of blockchain records ensures that the data provenance remains tamper-proof, further enhancing the trustworthiness of the LLM. Our approach leverages three critical features of BCT to strengthen LLM security: 1) Transaction Anonymity: While data provenance is recorded on the blockchain, identities of data contributors can be anonymised, protecting their privacy while ensuring data integrity. 2) Decentralised Repository: Enhances the system's resilience against potential attacks by distributing the data provenance record across the blockchain network. 3) Block Validation: Rigorous consensus mechanisms ensure the validity of each data point added to the LLM's data supply chain - minimising the risk of incorporating inaccurate or manipulated data into the training process. Using the experimental approach, initial evaluations using simulated LLM training data on a blockchain platform demonstrate the feasibility and effectiveness of the proposed approach in enhancing data integrity. This approach has far-reaching implications for ensuring the trustworthiness of LLMs in various applications.

Keywords: Blockchain-based data provenance, LLM security, Generative models

1. Introduction

The progression of Generative AI (GenAI) technology, particularly Large-Language Models (LLMs), has revolutionised many fields. From machine translation and content creation to chatbot development, LLMs have changed how we use AI and continue to shape the future of AI-powered tools. As such, the security and trustworthiness of LLMs are paramount. Untrustworthy LLM provenance, meaning a lack of transparency about the data used to train the model, can introduce biases, facilitate manipulation, spread propaganda, and create security vulnerabilities (Singh, 2024; Wang *et al.*, 2024). For instance, an LLM trained on biased data could perpetuate societal stereotypes in its outputs. Traditional methods for LLM securing supply chains may not be sufficient for the complexities of LLM development, highlighting the need for a more robust approach (Luo, Luo and Vasilakos, 2023; Wang *et al.*, 2024; Zuo *et al.*, 2024).

Blockchain technology (BCT), designed to provide integrity and safety within the fintech sector, provides a compelling case for robust security integration within LLMs. With its core principles of a distributed ledger, immutability, and secure consensus mechanisms, BCT offers a promising solution for enhancing LLM supply chain provenance, promoting decentralisation and immutability (Treiblmaier, 2018; Himeur *et al.*, 2022). By leveraging blockchain, we can create a transparent and tamper-proof record of every step in the LLM lifecycle, from data collection and model training to end-user interactions and fine-tuning (Luo, Luo and Vasilakos, 2023; Zuo *et al.*, 2024). We hypothesise/argue that this increased transparency can significantly improve trust and confidence in the integrity of LLMs.

Building upon our previous work on securing LLMs with zero-knowledge proofs, this research investigates the potential of BCT to enhance LLM supply chain provenance (Singh, 2024). Our research focuses on designing and conducting a series of simulations assessing a blockchain-based LLM's core functionalities. By leveraging existing LLM architectures and blockchain development techniques, we can create a simulated environment to explore enhanced security, scalability, and trustworthiness through transparency.

Through these simulations, we aim to address the following research questions:

RQ1: How can BCT be implemented to create a secure and verifiable record of data provenance throughout the entire LLM data supply chain?

RQ2: What specific data points within the LLM supply chain are most critical to track and verify using blockchain?

RQ3: How can BCT mitigate the risks of bias, manipulation, and security vulnerabilities introduced by untrustworthy LLM provenance?

RQ4: How can blockchain-based provenance systems be designed to ensure scalability and interoperability for large-scale LLM deployment?

RQ5: How can blockchain-based provenance systems be designed to promote transparency and trust among stakeholders involved in the LLM development process?

Our work aims to ensure verifiable and secure data provenance throughout the LLM lifecycle, drawing inspiration from the architectural solutions proposed by Zuo *et al*. (2023). By analysing the results of these simulations, we aim to contribute valuable insights into the feasibility and effectiveness of blockchain-based LLMs. This knowledge can inform future research and development efforts to create secure, trustworthy, and transparent LLMs empowered by blockchain technology.

This serves as a cornerstone paper outlining the trajectory for our research and future research endeavours on blockchain-based LLM security. As such, concerns and questions outlined in this paper may only partially be answered or addressed within the paper itself and will be the focus of our future work on the topic. It is essential to highlight the broad scope of the research at the outset to plan a road map of objectives we set out to achieve.

2. Literature Review

This literature review focuses on the emerging concept of LLM supply chain provenance and explores the potential of blockchain technology in addressing its challenges. Focusing on research published within the last eight years (2016-2024), this review aims to identify key themes in LLM supply chain research, analyse the potential of blockchain technology, and identify gaps in the research.

2.1 Trust in LLM Supply Chains

As LLMs become more prevalent, concerns regarding data bias, security vulnerabilities, and the interpretability of their outputs are paramount (Balayn *et al.*, 2024). Therefore, the need arises to create more robust LLM supply chains to facilitate this data transfer to and from the LLM on large-scale applications, such as Large Multimodal Model (LMM) applications. Defined as a simple three-phase circuit of events comprising LLM Infrastructure, Model Lifecycle, and Application Ecosystem, the LLM supply chain runs deep into the LLM training and data transfer processes where it monitors data flow from the infrastructure development level to the model lifecycle and finally to the end application ecosystem (Wang *et al.*, 2024).

This supply chain is responsible for the trust processes of the LLM, both human and non-human entities. As multiple organisations or individuals are likely to be involved in the LLM development, these trustor and trustee relationships shape the supply chain as producers, consumers, or indirect stakeholders (Balayn *et al.*, 2024). Should these supply chain actors not adhere to strict protocols regarding data cleaning and ethical and transparent operational processes, they could taint the LLM to adhere to biases which would affect the user of the LLM-powered application (Wang *et al.*, 2024).

Establishing trust in these systems, therefore, relies on verifiable data provenance and secure training processes (Balayn *et al.*, 2024; Wang *et al.*, 2024). Trustworthy AI requires explainability, fairness, and robustness of trust-based LLM solutions (Singh, 2024). Therefore, understanding the origin and characteristics of training data allows for identifying and mitigating potential biases, promoting fairer AI development. Secure training processes and data transfer are crucial to prevent malicious actors from manipulating data and compromising the integrity of AI models and their downstream applications (Luo, Luo and Vasilakos, 2023; Wang *et al.*, 2024).

2.2 Blockchain-Based Solutions for LLM Supply Chain

Invented in 2008 by Satoshi Nakamoto, Blockchain is a distributed, immutable ledger built for tracking transactions and providing an un-tamper-able, verifiable provenance record (Guo and Yu, 2022). Primarily used in fintech, this technology tracks the transactions of digital currency and cryptocurrency between entities and records these transactions in a block. Blocks are then verified by other users, called miners, through consensus mechanisms – a voting method where nodes within the network work to verify the transactions within the block (Himeur *et al.*, 2022). If transactions are valid, the block is added to the Blockchain and will be uneditable. Since the addition of new blocks to the chain relies on the hash values of the immediate previous block, any attempts to alter the chain will cause ripple effects breaking the transactions to follow, making it easy to identify where the tampering occurred (Guo and Yu, 2022; Himeur *et al.*, 2022).

BCT, therefore, presents a promising approach for establishing secure and transparent supply chains (Apte and Petrovsky, 2016; Treiblmaier, 2018) and can be a vital tool to secure LLM supply chains (Luo, Luo and Vasilakos, 2023; Zuo *et al.*, 2024). Luo, Luo and Vasilakos (2023) propose a framework called BC4LLM that leverages blockchain to ensure the security of the entire LLM training process, including creating a reliable learning corpus by verifying data ownership ensuring data security, and facilitating secure training processes. However, two drawbacks of this solution are reduced scalability and high energy consumption.

Similarly, Fan *et al.* (2023) introduced a federated learning framework called FATE-LLM, offering a promising approach for addressing data privacy and security concerns within the LLM supply chain. Federated learning allows for training LLMs on distributed datasets without requiring data to be transferred to a central location. This approach mitigates the risk of data breaches and empowers data owners to maintain control over their data. However, federated learning also presents challenges, including communication overhead between training clients and the central server.

2.3 Trusting and Scaling Blockchain-Based Approaches

Blockchain-based approaches to cloud and other distributed computing platforms see a lack of consumer confidence due to lack of transparency and loss of control over data (Gong and Navimipour, 2022). Further potential limiting factors to the adoption of Blockchain-based approaches include lack of degraded networking control, susceptibility to majority attacks, and irreversible commits to the blockchain (Zhu *et al.*, 2019). Gong and Navimipour (2022) highlights research suggesting possible solutions to these problems such as Controllable Blockchain Data Management which implements a Trust Authority node to defend against majority attacks (Zhu *et al.*, 2019), and scalability and reliability of blockchains assessed by Rimba *et al.* (2020) found that their cost model strengthened their outcome reliability and helped justify their costs.

Further, the Blockchain-enabled IoT approach explored by Singh, Rathore and Park (2020) explore AI-based strategies to minimise the computational overhead required for blockchain mining. The approach addresses the blockchain architecture and investigates the use of path discovery, data association, prediction, aggregation, and classification to improve the performance of blockchain miners, thus decreasing the mining cost and computational overhead required. Additionally, this integration can be effective in increasing the scalability of the solution, essentially fending off 51% majority attacks in the process (Singh, Rathore and Park, 2020). To further address the issue of scalability, the researchers again approach the problem using federated learning to distribute the sheer capability of the system as opposed to opting for a centralised approach to data storage.

The literature addresses approaches to facilitate LLM secure training processes, mitigate the risk of data breaches, and highlight the potential for scalability challenges of Blockchain-based LLMs. However, further research is required to understand these frameworks' practical applications and limitations regarding scalability, adoption, and field testing. Additionally, research thus far has yet to address the challenge of how untrustworthy actors can be identified and dealt with from within the Blockchain itself. The verification and validation of sources rely on the source creator's trust, and the accuracy of the information cannot be validated before it is used in the training process. No framework highlighted the explicit dataset an LLM utilised to generate its response. These are all issues we aim to address in this research.

3. Methodology

This chapter outlines the methodology employed in our research to explore Blockchain-based LLMs' potential. As our focus lies in understanding the practical functionalities of this technology, we adopted an empirical approach centred on simulation and evaluation.

3.1 Research Objectives

Design and conduct simulations that evaluate the functionalities of a Blockchain-based LLM to:

- Investigate the feasibility of utilising BCT to create a verifiable and secure record of data provenance throughout the LLM lifecycle.
- Explore how Blockchain can mitigate the risks associated with untrustworthy LLM provenance, such as bias and manipulation.
- Analyse the potential of blockchain-based LLMs to enhance security, scalability, and trust within the LLM development process.

3.2 Simulation Design

3.2.1 Core components

Our simulations centred on two core components:

- **LLM Architecture:** Drawing inspiration from open-source libraries to influence our design, we achieved our objective of simulating core LLM functionalities, leveraging a wide variety of open-source LLMs and allowing us to choose custom embeddings and vector databases (obtained from HuggingFace).
- **Blockchain Development:** We developed a custom-simulated blockchain architecture to monitor the various components of the supply chain using established blockchain development techniques.

Simulation Scenarios:

Simulations were designed to assess the provenance tracking ability of blockchain-based LLMs and the database detection and mitigation strategies that they offer. Simulations scenarios were grouped into three categories:

- Simulation 1: **Verifiable data provenance**: This scenario simulated data flow throughout the LLM supply chain. Relevant data points such as model name, embeddings, vector database utilised, etc., were logged as entries on the blockchain. This provides a verifiable chain of events leading to response generation.
- Simulation 2: **Bias mitigation**: This scenario explored the bias mitigation techniques implemented within the blockchain-based LLM supply chain. Analysis of sources and data categorisation are critical components of this simulation.
- Simulation 3: **Data manipulation and Security Analysis**: This scenario focused on simulating security breaches and unauthorised access attempts. We evaluated the effectiveness of blockchain's security features in preventing unauthorised data manipulation and protecting the integrity of the LLM.

3.3 Data Collection and Curation

Since this research focuses on assessing the viability of Blockchain-based LLMs rather than on the overall training and fine-tuning process, we have chosen to opt for a custom dataset comprising 200 academic research papers. This allows us to effectively construct simulations of adding and modifying source data and evaluate Blockchain security's effectiveness. Thus, this requires more than a few static datasets and must involve a multitude of text-based data with verifiable sources, even if the total size of the data is smaller.

Data on the simulation design of Blockchain-based LLMs were collected by consolidating information from the literature review with our series of trial-and-error designs. Each design was evaluated and contributed towards the final simulation on which we base our findings.

4. Findings

This chapter explores the implementation of BCT to create a secure and verifiable record of data provenance throughout the entire LLM lifecycle.

Simulation 1: Verifiable data provenance and crucial supply chain data points

Blockchain design and security architecture (RQ1)

Our blockchain architecture focused on modularity and utilised three chain types: Data Vendor chain, Data Categorisation chains, and Master chain. Each block comprises the following standard components: block hash, block header, previous block hash, Merkle root, nonce value, and timestamp. Our design leverages the permissioned Blockchain architecture, requiring all users to provide credentials before accessing any functionality. User credentials and every other data point monitored by the blockchain utilise the secure one-way cryptographic hashing algorithm, SHA256, to ensure anonymity whilst preserving the data in a tamper-proof format. Any alterations to these data points will generate a new hash, prohibiting changes once the data point is logged and passed to the blockchain.

Leveraging multithreading to represent worker nodes on our simulated distributed network, we stipulate that nodes must evaluate the chain's current state before any new data can be written to the blockchain. This ensures uniformity of the chain state and protects against chain corruption. Data points representing interactions with the LLM are then stored on a block; these interactions vary depending on the chain – Master, Data Vendor, or Data Categorisation. Unique data points for each chain are listed in Table 1 below. These blocks are mined by

the network nodes utilising the Proof-of-Work consensus mechanism. Once a block is successfully mined, it is added to its respective chain, where it lives in an un-tamper-able state, as any alterations to this entry will derail future transactions.

Blockchain and the LLM supply chain (RQ1 & RQ2)

Our design builds on the data collected from the literature and specifies unique data points that represent the individual LLM-specific aspects. Table 1 highlights the data points for the three chains used in this version of our simulation.

Table 1: Shows the different chain types and which data points each chain monitors

UNIQUE DATA POINTS	
MASTER CHAIN	Model
	Embeddings
	Vector database
	Vendor
DATA CHAIN	Category
VENDOR CHAIN	Vendor
DATA CATEGORISATION CHAIN	Category

The Blockchain tracks LLM source data upload through the Data Vendor and relevant Data Categorisation chains. The Data Vendor chain tracks the entity that supplied the data and the specific Data Categorisation chain logs where the data is stored. When more data are added to any chain, the same mining principles apply, and a previous state cannot be reverted to or altered without affecting succeeding blocks. Only once all blocks related to data supply are successfully mined, the data get added to the vector database for the LLM to consult when answering user prompts.

When a user interacts with the LLM, the chat interaction gets logged on the Master chain. Data such as model used, embeddings, chat history, vector database, temperature, data vendor, source data category, and user wallet ID are captured within a block.

Simulation 2: Bias Mitigation (RQ3)

This simulation assesses Blockchain's ability to identify and prevent potential biases that supply chain actors have on LLM outcomes. This research defines bias as any overreliance or undue influence on a particular outcome. To achieve this, we set out the following preconditions we believe influence potentially biased outcomes:

1. LLM's overreliance on specific data sources
2. Source database lacks sufficient data on a topic
3. Source database does not include data from enough vendors on topics
4. References provided by LLM could not be validated
5. High temperature settings of the LLM

This simulation leverages the blockchain architecture pinning it against the stock LLM to determine whether the Vendor and Data Categorisation chains benefit the LLM in bias mitigation. This simulation was conducted with these chains not active, *Case 1*, and active, *Case 2*, to determine a before and after result. Prompts and datasets were kept constant, and temperature settings were incrementally increased by 0.1 degrees, starting at a temperature of 0.1 until a peak of 1.0.

Table 2: Shows the average of results from 200 total prompts

	Responses Containing Bias /Per 10 Prompts	Unverified Sources /Per 10 Prompts	Repeat Sources /Per 10 Prompts	Total Sources Consulted /Per 10 Prompts
Case 1 – Chains Not Active	3.46	1.1	2.36	4.16
Case 2 – Chains Active	1.42	0.2	1.22	4.86

As shown in table 2 above, from the 200 prompts inputted, Case 1 responses were particularly susceptible to influential bias, with an average of 3.46 responses containing bias, per 10 prompts. The most significant contributor to this metric was referencing sources that could not be verified at a rate of 1.1 unverified sources for every ten responses. The LLM also displayed an overreliance on a handful of sources, listing an average of 2.36 repeat sources out of 4.16 total sources consulted, per 10 responses.

Case 2 generated contrasting results as we observed the influential bias of responses decrease to an average of 1.42 responses containing bias, per 10 prompts. Unverified sources were reduced to 0.2, per 10 prompts, and repeat sources were, on average, 1.22 out of 4.86 total sources consulted, per 10 responses.

Simulation 3: Data Manipulation and Security Analysis (RQ3 & RQ1)

Table 3 details desired and actual outcomes of the security analysis by stipulating the alterations made to the blockchain and their potential implications. The ideal outcome of this experiment is to have a score of zero, as any higher would indicate a potential vulnerability.

Table 3: Shows outcomes of security analysis testing conducted

Types of alterations	Proposed effect on the Blockchain (desired outcome)	How it affected the Blockchain (actual outcome)	Implication weight	Actual Weight
Block Hash	Disrupt the flow of data between user-LLM into the blockchain and alter new and existing transactions	Every instance of tampering related to the block hash manipulation was successfully detected and stopped	5	0
Nonce	Attempt to alter a new transaction not yet mined to represent and change properties not related to user-LLM interaction	No alteration of transactions was successfully added to the blockchain	5	0
Source Dataset	Infect the source information and introduce bias/misinformation into the LLM's response computation	Any changes to source data were detected before the LLM execution began	5	0
Timestamp	Lesser attempt but an indication that there were efforts to sabotage the blockchain	No alteration of timestamp was recorded on the blockchain	2	0
New (malicious) data source	Attempt to introduce a malicious actor to supply information	The system allowed the creation of a new actor	3	3
51% attack	Attempt to overwhelm the blockchain by staking a claim that majority of the nodes agree with the altered state of transactions as opposed to the original	This typical attack was neutralised by the simulation keeping the blockchain state in memory	4	0

From this table, we can detect one source of potential vulnerability with the blockchain pertaining to source creation.

5. Discussion

This chapter delves into the implications of the results presented, exploring the potential of BCT for establishing a secure and verifiable record of data provenance throughout the LLM lifecycle.

These findings hold significant value for the development of secure and trustworthy LLMs. Establishing a verifiable record of data provenance allows stakeholders to gain insights into the data used to train LLMs,

fostering transparency and accountability within the LLM development process. By mitigating risks like manipulation and bias, BCT can contribute to producing more reliable and unbiased LLM outputs.

5.1 Evaluation of Findings

Research Question 1: Implementing BCT for Secure and Verifiable Provenance Tracking

To assess the verifiable record of data provenance, we adopted an object-oriented software approach to our blockchain design. Leveraging the concept of abstraction, we designed the blockchain to accommodate various individual components that symbolise the flow of data. This flow of data is monitored on a master chain, whilst smaller individual chains handle the responsibility of security and integrity of data sources and the categorisation of data for easily monitoring any amendments to the data and source catalogues of the LLM.

Ensuring a verifiable record of transactions requires a clear understanding of where each transaction is going, meaning what task is carried out by executing that transaction. We have broken down the transaction process into three streams of data flow: an input stream of data, a verification stream, and a chat stream.

The data input stream begins with vendor validation, as identified within the literature review; data streams may not be from the same source. Data quality is crucial in ensuring accurate and reliable LLM computations, and vendor validation is essential to ensure accountability for the data supplied to the LLM.

Research Question 2: Tracking and verifying critical supply chain data

Our simulation design evolved as we conducted this research, and more demand was placed on the blockchain. We initially identified the blockchain components based on the findings within the literature and concluded that a more robust solution was required. We created a verifiable architecture for our final implementation using three blockchains: Vendor, Data Categorisation, and Master chain. The standard and unique data points identified for use within these chains provide crucial provenance tracking capabilities that allowed for the use of this design in answering other questions highlighted in this research.

Although further research is required, this approach to architectural design provides a solid foundation for scalable solutions as the demands of the blockchain increase and applications require domain-specific knowledge.

Research Question 3: Mitigating Risks with BCT

Simulation 2 allowed us to focus on the ability of the blockchain architecture to provide meaningful contributions to bias mitigation within the LLM response generation process. As per the definition of bias described above, bias in the context of LLMs can stem from overreliance on specific data points and sources to answer a particular type of prompt. As such, skewed or unjust data providers can input false data if not verified.

By ensuring the LLM consults more than one source of information before generating its response, we can limit the potential for bias to creep into the LLM's output should the consulted data be skewed in any manner. This is especially useful in scenarios of potential vulnerability where the data vendor is valid, but the source data is only partially bias-free.

Simulation 2 shows that BCT's architecture provides a solid foundation for data vendors to provide legitimate and accurate data and reduces the challenges associated with interdependencies of, and interactions with the various supply chain actors (Wang *et al.*, 2024). Thus, this aids in prohibiting unwanted bias/influence on the LLMs response generation and provides a more significant deal of trust that the data and output will be trustworthy.

Simulation 3 demonstrated the potential of BCT to mitigate security risk manipulation. The immutability of the blockchain ledger prevented most attacks relating to data manipulation however further investigation is required on how to best prevent malicious data vendors from entering the network.

Research Question 4: Scalability and Interoperability for large-scale LLM development

Although not explored experimentally in this work, as BCT platforms already host fair degrees of relative scalability, were this limit reached in the context of considerably large LLMs, we can see a strong argument for the innovation of hybrid blockchain technology - utilising low-computation-cost blockchains to supplement a larger blockchain where expansion of capacity would be required.

Research Question 5: Promote transparency among stakeholders

Although not explored experimentally, our current research suggests the use of transparency-chains to monitor and ensure that each occurred transaction abides by the regulations stipulated by an organisation. Blockchain-based LLM solutions can therefore be tailored to meet organisational needs.

5.2 Limitations of This Study

Blockchain assumes that data added to the chain is correct and final. Thus, Blockchain does not allow any amendments to existing data on the chain. Therefore, this study did not take the action of updating or removing chain-data into consideration. Further, the current representation depicts that we have a solution for an immutable supply chain – essentially a snapshot of the chain, as the chain must alert us when a change occurs. However, the supply chain is not stagnant – it grows and changes with the addition of new data sources. Future work will consider expanding this chain to include elements, such as web supplementation, where every time the LLM consults a new source, it creates a new chain.

5.3 Recommendations and Future Work

This research lays the groundwork for further exploration of BCT in LLM provenance tracking. Future directions include:

- Scalable Permissioned Blockchain Architecture: Investigating and implementing BFT protocols to enable a robust and scalable network for real-world LLM training scenarios.
- Enhanced Data Validation Techniques: Integrating machine learning models for data quality assessment and anomaly detection within the validator script.
- Integration with LLM Development Tools: Developing APIs to facilitate seamless integration of the BCT system with existing LLM development workflows.
- Smart Contract Implementation: Exploring the development of smart contracts to automate key functionalities like data validation, model lineage recording, and access control on the blockchain.

These future research directions promise a more secure, transparent, and trustworthy LLM development ecosystem. By establishing a verifiable record of data provenance, BCT empowers stakeholders to assess potential biases, identify manipulation attempts, and ultimately foster trust in the outputs generated by LLMs. This, in turn, can pave the way for the responsible and ethical development and deployment of LLMs, unlocking their full potential to benefit society.

6. Conclusion

As GenAI advancements become increasingly complex, the necessity for a robust foundation in cybersecurity becomes an ever-growing challenge (Singh, 2024). LLMs are the cornerstone of many GenAI applications, acting as the engines that translate instructions from human users into tangible outputs. Therefore, securing the LLM supply chain – the entire process from data collection to model deployment – is paramount for ensuring GenAI's trustworthiness and ethical use.

Previous research has explored promising avenues for LLM security, with zero-knowledge proofs demonstrating potential in certain aspects. This work builds upon this foundation by investigating the potential of blockchain technology to enhance LLM supply chain security further. Our core hypothesis centred on the idea that blockchain's core principles – immutability, secure record-keeping, and verifiable transactions – could solve the challenge of untrustworthy LLM provenance.

The results of these simulations offer valuable insights into the feasibility and effectiveness of blockchain-based LLMs. They demonstrate the potential for blockchain technology to create a secure and verifiable record of data provenance throughout the LLM lifecycle, addressing the core research questions outlined at the outset. This research paves the way for architectural solutions integrating blockchain to ensure secure and verifiable data provenance within LLM supply chains. Building upon the groundwork laid by Luo, Luo and Vasilakos (2023) and Zuo *et al.* (2024), future research can delve deeper into specific architectural designs and explore their practical implementation. Ultimately, our work contributes to creating secure, trustworthy, and transparent LLMs empowered by blockchain technology.

Acknowledgement

The authors would like to acknowledge the financial contribution made by Mr Luke Vorster from the School of Mathematics, Statistics, and Computer Science at The University of KwaZulu-Natal, South Africa, for funding the

publication and presentation of this research at the 4th International Conference on Artificial Intelligence Research 2024.

References

Apte, S. and Petrovsky, N. (2016) 'Will blockchain technology revolutionise excipient supply chain management?', *Journal of Excipients and Food Chemicals*, 7(3), pp. 76–78.

Balayn, A., Yurrita, M., Rancourt, F., Casati, F. and Gadiraju, U. (2024) 'An Empirical Exploration of Trust Dynamics in LLM Supply Chains'.

Fan, T., Kang, Y., Ma, G., Chen, W., Wei, W., Fan, L. and Yang, Q. (2023) 'FATE-LLM: A Industrial Grade Federated Learning Framework for Large Language Models'. arXiv.

Gong, J. and Navimipour, N.J. (2022) 'An in-depth and systematic literature review on the blockchain-based approaches for cloud computing', *Cluster Computing*, 25(1), pp. 383–400.

Guo, H. and Yu, X. (2022) 'A survey on blockchain technology and its security', *Blockchain: Research and Applications*, 3(2), p. 100067.

Himeur, Y., Sayed, A., Alsalemi, A., Bensaali, F., Amira, A., Varlamis, I., Eirinaki, M., Sardianos, C. and Dimitrakopoulos, G. (2022) 'Blockchain-based recommender systems: Applications, challenges and future opportunities', *Computer Science Review*, 43, p. 100439.

Luo, H., Luo, J. and Vasilakos, A.V. (2023) 'BC4LLM: Trusted Artificial Intelligence When Blockchain Meets Large Language Models'.

Rimba, P., Tran, A.B., Weber, I., Staples, M., Ponomarev, A. and Xu, X. (2020) 'Quantifying the Cost of Distrust: Comparing Blockchain and Cloud Services for Business Process Execution', *Information Systems Frontiers*, 22(2), pp. 489–507.

Singh, S. (2024) 'Enhancing Privacy and Security in Large-Language Models: A Zero-Knowledge Proof Approach', *International Conference on Cyber Warfare and Security*, 19(1), pp. 574–582.

Singh, S.K., Rathore, S. and Park, J.H. (2020) 'BlockIoTIntelligence: A Blockchain-enabled Intelligent IoT Architecture with Artificial Intelligence', *Future Generation Computer Systems*, 110, pp. 721–743.

Treiblmaier, H. (2018) 'The impact of the blockchain on the supply chain: a theory-based research framework and a call for action', *Supply Chain Management: An International Journal*, 23(6), pp. 545–559. Available at: https://doi.org/10.1108/SCM-01-2018-0029.

Wang, S., Zhao, Y., Hou, X. and Wang, H. (2024) 'Large Language Model Supply Chain: A Research Agenda'.

Zhu, L., Wu, Y., Gai, K. and Choo, K.-K.R. (2019) 'Controllable and trustworthy blockchain-based cloud data management', *Future Generation Computer Systems*, 91, pp. 527–535.

Zuo, X., Wang, M., Zhu, T., Zhang, L., Ye, D., Yu, S. and Zhou, W. (2024) 'Federated TrustChain: Blockchain-Enhanced LLM Training and Unlearning'.

Generative AI and the End of Education

Caroline Stockman
University of Winchester, UK

caroline.stockman@winchester.ac.uk

Abstract: This paper reviews the prevalent rise of generative artificial intelligence and its impact on HE in the UK. In doing so, it echoes a central thought from Neil Postman's provocatively named book which reflects on the nature and purposes of education. Whilst GenAI may be approached with a constructivist disposition, much of the response in practice and policy is marked by technological determinism. A constructivist approach, however, allows us to (re)consider the nature and value of education. The paper therefore maintains the angle of the educator in asking: what are we trying to do in (higher) education? What is the purpose of Higher Education today? This is a cross-disciplinary question, much like Postman's propositions were a cross-curricular reflection on the nature of schooling. Postman looked for a unifying narrative that can inspire 'the ends of education', or what education actually is and tries to accomplish, before considering its tools and approaches towards those goals. Those ends ensure that education does not become subjected to the false gods, such as economic utility and consumerism. These would spell the demise of any meaningful education, or 'the end of education'. What appears foundational to these questions is a belief in the nature of human potential. In education, the cultivation of that potential is arguably a fundamental end, whichever way education is organised. Where the cultivation of that potential leads, however, can remain rooted in a humanist framework, it could become posthumanist, or it could be simply bewildering. With the means of generative AI, the analysis raises questions on epistemic threat versus intellectual success, and new horizons of creative possibility in human-computer interaction. The human potentiality of thinking, interpreting, criticality, and scepticism, come forward as retained elements of the humanist narrative. In the dawn of generative AI as a teaching, learning, and assessment instrument, the end of education remains human potential. Higher Education especially, remains the place to critically and ethically stimulate the human mind towards new horizons of knowing.

Keywords: GenAI, Purpose of education, HE, Technological determinism, Postman

1. Introduction

AI may be perceived to be a single technology, yet it is 'a jungle ecosystem' of techniques, theories, and values (Lea, 2020, p323). It may be perceived as new, yet – much like many digital technologies – it has a long history of human thought and practical invention which has infused it with meaning. At present, there is a sense of accelerated change. On generative AI (GenAI), the UK government and advisory bodies have issued various national guidance and position statements in the last two to three years. These apply to education generally, Higher Education (HE) more specifically, and to wider society. Much of the current hype has emerged from ChatGPT, an OpenAI product. Though this paper isn't exclusively focused on ChatGPT, it illustrates the type of technology which spurred policy makers and practitioners to respond. Text-based GenAI in particular has made a developmental leap, as part of a 'GenAI boom' according to *MIT News* (Zewe, 2023). GenAI are technologies which 'generate' text, audio, images, and other content. This content may be considered as new or original, rather than a reproduction of existing content. An algorithm will be 'trained' on existing content that relates to its anticipated functions of text, image, sound, or other media generation. With predictive modeling, it will 'learn' to recognise and reproduce patterns in the existing content based on human prompting.

Two central philosophies of technology, being technological determinism or its opposite of social constructivism, offer a useful lens to analyse the responses to GenAI in policy and practice. This will be the first section below. The paper will, secondly, apply this to policy-making for UK HE. What will emerge is that 'how' we approach HE should temporarily take the backseat to a fundamental question: 'why'. This introduces the perspective of Neil Postman, 20th century American theorist, media critic, and educator. In his book *The End of Education* (1996), Postman expresses his view that any conversation on education is often about means, rather than ends. By this he means education would benefit from having one overarching purpose that steers policy-making and pedagogies. So amidst uncertain futures, is it possible to distinguish one such end of HE?

2. Technology Determinism and Constructivism

When GenAI is deemed 'unstoppable whatever government leaders do' (Crawford, 2024), or 'poised to change society more than any technology before it' (McGregor, 2024), this implies a determinist lens. Technological determinism attributes a great deal of causal power to technology's impact on human society. This may have a positive outlook, synonymising technological development such as GenAI with human development, as stepping stones of progress. In (Higher) education, that may offer the perspective that 'technology for teaching will *in and of itself* lead to enhanced or transformed educational practices' (original emphasis in Kirkwood,

2015). On the other hand, it may take a more dystopian view: things will change, but for the worse. Wyatt (2008) discusses nuances, from radical causation to more dynamic sociotechnical ordering, where technology nonetheless has a central influence still. So while it is a deceptively simple philosophy, it still functions as an umbrella concept (Hallström, 2020), and some may say to adopt a 'soft' technological determinist perspective (Kirkwood, 2015). Similarly, as discussed by Hallström (2020), there are streams of technological determinism which consider technology to be a neutral force, and those that recognise human values and processes. Nevertheless, in the continuum of theoretical nuance, determinism still posits technological 'impact' as a central thought.

A common critique on (radical) technological determinism is that it leaves no space for human free will, autonomy, or agency. Constructivist perspectives, on the other hand, emphasise human ability to actively steer and shape change. This human ability has been posed as a human right (freedom of choice), and related to human dignity, responsibility, and other values. Within constructivist philosophies of technology, there is still a spectrum of theorising. Some models like ANT (Actor-Network Theory) suggest that the driving force of change is situated in a network of social processes, rather than with any particular entity such as technology. Change occurs holistically in the coming together of several parts – people, technologies, but also texts, contexts, processes, etc. A similar proposition is SCOT, or the Social Construction of Technology. This model focuses on the social processes that originated a technology in the first place. It therefore denies the possibility of technology as a neutral, outside force, by placing people and social processes as direct influencers of technology's emergence and also later uses.

Words like 'artificial' and 'intelligence' are value-laden, and subject to hyperbole and over-selling (Selwyn, 2022). Yet terminologies such as intelligence, or machine 'learning', are educationally significant. The denotations show a constructivist perspective, either accidental or intentional. It is complicated further by various interpretations on the nature of 'education', 'teaching', or 'intelligence' in itself, before considering what 'technology' or 'AI' is in that mix. The 'intelligence' of GenAI is ultimately the production of new content based on a human prompt. Interestingly, this technical aspect has been the foundation of a critique on GenAI as being inherently deterministic. While it may appear to generate new content, it is always based on a calculation and re-ordering of old data. That means something which is deterministic, such as the functioning of a machine, may still be experienced as constructivist, especially when it mimicks human authenticity, harkening back to the Turing test, and historically lingering questions on the locus of consciousness.

For now, it is important to keep recognising GenAI 's core functioning as determinist, meaning its output has a certainty based on the underlying calculations and data. Yet the human hand in this construction cannot be denied either. The 'technology complex' (Fleck & Howells, 2001) consists of physical artefacts, scientific thought, software processes, industry structures, management techniques, social and legal frameworks, and more. Impacts may indeed arise from any of these elements, they may be certain and predictable. Fundamentally, the underlying model has been built by people over time, in a particular place and context, which infuses it with beliefs and values that may differ from another time and place. The choice of data, for example, on which the GenAI will be trained, has been found highly significant to its outputs. Then, it very much depends on what we, as people, do with it. It is easy to prompt GenAI in such a way that it produces output which is wrong, or biased one political way over another, or socially or emotionally slanted. What we choose to do next with this output, is another constructivist exercise. It is not certain – it depends on our choices, our thinking. This re-emphasises the need to examine our social processes that integrate GenAI such as education.

Technological determinism cannot be dismissed as a fallac, shown by its persistence in research, public opinion, practices and policies (Wyatt, 2023). Yet Oliver (2011) cautions that we should be 'avoiding simplistic claims about impact, effect and technical causation, and concentrating instead on descriptions of practice, accounts of purposeful action and negotiated meanings' (p.382). This relates to another critique on technology determinism, that it seemingly absolves people of responsibility for changes or consequences. The impact is certain and unavoidable, after all. Again, a rather dangerous postulation. As argued above, we must continue to approach GenAI (or any technology, for that matter) as a constructivist exercise. The next section explores what the UK's national statements on (generative) AI have revealed in that regard, and HE's dominant response in return.

3. GenAI Meets UK HE

A constructivist analysis of technology must attribute central importance to technology in a particular time and place. In what follows, the context of HE in the UK will be examined. HE is generally understood to mean

undergraduate and postgraduate degrees offered to students above the age of eighteen. HE institutions in the UK can operate with a high degree of organisational independence, confined within national policy and strategising. Though there are around 450 HE providers in the UK, over 400 are in England. The majority of these do not have their own 'degree-awarding powers'. Instead, they offer the teaching of degrees which will then be awarded by another institution once requirements have been fulfilled. There are 175 such 'recognised bodies' in the UK which have the right to award degrees.

At the end of 2023, the UK's Department for Education published a position statement on the use of GenAI in the education sector (DfE, 2023). It aligns with the UK government's broader disposition towards greater technology acceptance and development, highlighted by the pro-innovation approach to AI regulation (Gov UK, 2021; 2023), the pro-growth National Data Strategy (Gov UK, 2022), and the government's educational technology strategy (DfE, 2019). It is constructivist in recognising there is potential but also limitation and possible harm depending on the use of GenAI. However, it aligns with a societal sense that GenAI is unavoidable and impactful. Its position is that education 'needs' to act in response to the technological change, which firstly includes an emphasis on building long-term knowledge to enable human subject expertise, for which the AI is not a substitute. Secondly, it suggests a revision of skills training is needed to adequately prepare for changing workplaces, and the safe but effective use of GenAI. This early position statement was followed up by a summary of selected educator and expert views in January 2024 (DfE, 2024). This report echoes the sense of technology's transformative impact on education. It suggests GenAI is already used across most of the school curriculum, to aid language learning, for coding lessons, as maths coaches, for the creation of historical characters, or creative art, and more. Similarly, in tertiary education (e.g. further and HE), the advisory body JISC (2023) reports that GenAI is already used by students in various ways: as a writing aid, a study tool, a research and search platform, a creative art and visual production tool, and for maths and coding assistance. In addition, JISC (2024) also found uses for enhanced scheduling and task management, personal and emotional support. Much of this reporting focuses on 'how' it used, rather than the overarching 'why' of using it this way, another way, or not at all.

While HE does not have a national curriculum, it does have subject benchmark statements suggested by the QAA (the Quality Assurance Agency for HE). It is an independent, national body, which aims to maintain and enhance HE teaching and learning. The subject benchmarks set out the sector-wide reference points on the nature of study and graduate academic standards. Each statement is reviewed every seven years by a panel of UK subject specialists. In its latest cycle of benchmark reviews, GenAI was specified as one of the five subthemes for the new benchmarks, considering 'how' (rather than 'whether' or 'why') this technology should be used in different subject areas. In response to the government's position (DfE, 2023), the QAA emphasised the importance of critical AI skills to meet the needs of an evolving labour market, and the need for education to evolve in light of these technological developments (QAA, 2023). In a later publication, it describes the transformation of HE through AI as 'inevitable' (QAA, 2024, p.1), implying a determinist stance.

In line with national guidance, HE providers published their own institutional strategies. The Russell Group (an association of twenty-four high-ranking UK universities) published a set of five foundational principles in July 2023, which includes the commitment to support staff and students' AI literacy, and the need for staff to support students to use GenAI appropriately and effectively. It further states that universities 'will' adapt teaching and assessment, continue to ensure academic integrity and rigour, and share best practice for education. In relation to the second principle, student use of GenAI, the document acknowledges this will be subject-specific and adaptive to future technological development and work contexts. Overall, the document steers education to ensure effective use of GenAI.

More specific guidance has since been issued by individual institutions, such as the University of Oxford's guidance for students (2024) and King's College (2024). It states AI cannot replace human critical thinking, argumentation, or subject knowledge, but such technologies may be used in support of human growth and development. Other university alliances, such as GuildHE, the University Alliance, or the Cathedrals Group have not (externally) published group statements, but institutional strategies echo HE's dominant position to accept positive potential, with caution. The University of Winchester's (2024) guidance for students, for example, details that AI may be used to support learning by simplifying language or explaining complex ideas that might aid understanding.

There is also a common concern over academic integrity, particularly related to assessment (QAA, 2024). In a UK parliament research briefing, Felix & Webb (2024) alongside others (JISC, 2024) identified concerns over technology's impact on intellectual development, due to over-reliance on AI. A proliferation of teaching and

assessment strategies have been suggested to constructively mediate such issues, such as the 'Oxbridge tutorial method' (Wooldridge, 2023), UCL's three-tiered categorisation of AI use in assessment (UCL, 2024), or QAA's recommendation for critical pedagogy in HE (Seal & Smith, 2021).

Overall, a determinist perspective initially comes forward, in that GenAI is considered to have an unavoidable impact. That is significant, because it constitutes a collective human perception that this technology marks a certain change for human practice. But there is also a recognition of uncertainty. The impact or change seems tangible and certain, but what those changes entail is not yet (perceived to be) determined. That opens up a space for constructivist thinking. It shows a profound perception that human practices in teaching, learning, and assessment must adapt to a technological change, but also a belief in human potential to shape this change. The changes may incur positive potential, or harm, but it still depends on human use. Current strategising therefore appears to centre on 'how' to use and integrate GenAI in HE. However, a central question which still spans across these approaches is the overall purpose of HE. Rather than 'how' HE organises teaching, learning, and assessment, by using AI or not, it is still most important to ask 'why'.

4. False Ends of Education

In his book *The End of Education* (1996), Postman expresses his view that any conversation on education is often about means, rather than ends. He writes in the context of American public schooling of the 1990s, from his point of view as academic and HE tutor, and prolific media critic. He theorises that educational practice and policy remains stuck on the question of 'how' to integrate new technologies in teaching, learning, or assessment, instead of asking 'why'. Asking 'why' requires interrogating what the leading goals or values are. It requires the formulation of an overarching reasoning for all decisions and activities. Only after this, a discussion on means (the 'how') becomes meaningful. What Postman (1996) therefore proposes is a decidedly constructivist exercise, regarding the purpose of education as well as its means to achieve this purpose.

To Postman (1996), a lack of a clearly formulated purpose is an issue for education. He draws on religious terminology to explain his stance: 'For school to make sense, the young, their parents, and their teachers must have a god to serve, or, even better, several gods' (p. 4). That could be 'God' with a capital, but it doesn't have to be. What Postman values is for education to have some kind of transcendent idea to give a clear purpose, one that 'constructs ideals, prescribes rules of conduct, provides a source of authority, and, above all, gives a sense of continuity and purpose' (p.6). So the pursuit of God gives rise to certain approaches and practices in light of the overall goal. It ties people together in the same framework. This idea is clearly articulated as the goal, and while it may still have unknown qualities, it does create cohesion for the everyday decision-making. Actions and decisions are no longer random – they make sense, and that sense is derived from the overall goal. Without formulating such an end to education, it becomes meaningless (and therefore education ends, as the word play in the title of his book suggests).

The suggestions of technological determinism we can find in UK HE policy-making is alarming from Postman's (1996) framework. He theorises there are also 'false gods', and technology is one of them. It resonates throughout his other writings, such as *Technopoly: the surrender of culture to technology* (1992). He takes issue with changes that occur as a result of a view on technology as some kind of external, uncontrollable force (Postman, 1998). The UK Public Policy Lead at the Ada Lovelace Institute writes that 'the shape, pace, and direction of AI development in the UK is dictated not in Westminster or Whitehall, but overwhelmingly in the boardrooms and pitch decks of Silicon Valley.' (Davies, 2024). It is therefore constructed by people, but still certain in its impact – aligning with a 'soft' technological determinism. The rapid proliferation of policy and position statements that imply acceptance and change in HE learning, teaching, and assessment, purely as a result of GenAI becoming widely available, is indeed notable. It points to the central issue that the 'how' of education is driven by a wrong 'why'. That does not mean the prevalence of GenAI should be ignored altogether, which would undoubtedly do HE students a disservice. However, it does not justify GenAI to become the transcendent idea which leads HE (Postman, 1996).

A second false god which Postman (1996) identifies relates to the role of employability as an end of HE. Many students indeed pursue HE in preparation for certain careers (Brown, 2023). Though GenAI is positioned as impactful on human work and productivity, there is still uncertainty: how will GenAI redesign healthcare and nursing, or journalism, editing, copywriting, or website design and coding? That is a determinist question to ask, as we assume an impact will manifest itself, though we cannot say for sure what that will be. Yet taking human productivity and employment as the end of education is also false, in Postman's terms. He specified end values of economic utility and consumerism as problematic. To teach young people with the sole purpose of becoming economically useful – in generating money and then spending that money – would be a false god.

Postman did not deny the instrumental nature of schools altogether, but if that purpose was solely workforce preparation, that would be an issue: 'schooling can teach you how to make a living, or teach you how to live' (Postman, 1996, p.x). The role of the employability agenda in HE has been critiqued more widely, stating 'HE has been re-positioned as a means to an end rather than an end in itself' (Tight, 2023, np). It reduces the value of humanity by conceiving of education's primary aim as the production of good workers and consumers (Postman, 1996, p.31).

So it may be that learning with GenAI is meaningful, and therefore education must adapt its pedagogies to ensure this, but not if that learning is actualised to ensure the learner's employability (as an exclusive goal), or to accommodate technology. However, in the proliferation of policy and position statements, discussed in Section 4, appears a dual, constructivist stance. The impact of GenAI is perceived as certain, but the nature of its impact may still depend on human use. Concerns over the negative impacts on human thinking, integrity, and intellectual development, as well as hopes for its positive impact on the development of human learning and creativity, all share on central interest: human potential. This may lay the foundation to formulate HE's unifying narrative.

5. Human Retained

There has been a historic purpose of education to cultivate human potential, whichever guise education takes. 'Formation' is key, meaning education is an intentional process that shapes an underlying potential, which may be lost or corrupted. This intent to pursue education as humanisation is not unusual: 'For Kant, as it was for Plato and Rousseau, the "human" is the end of education' (Snaza, 2013, p.40). For as long as educationalists have set out their views on pedagogies or theories of the mind, there has been an unwavering belief in human potential. Piaget, Vygotsky, Malaguzzi, Socrates, Bruner, Wollstonecraft, Leopold, Mead, Noddings, Greene, and many more, all lived and worked in very different times and places but maintained a view that education must steer its means towards the cultivation of human potential. Without this purpose, education is nothing. It ends.

Human potential is the unifying narrative that appears to cross time and place of educational theory. The cultivation and protection of human potential could therefore be a viable end of education, as it resonates with educationalists across centuries. In HE, John Henry Newman's lectures captured in *The Idea of a University* are still considered foundational. Truth and the pursuit of knowledge should be overarching purposes, Newman stated, which can be achieved by debate, critical discussion and reflection, in multiple fields of study. Similar views are expressed in modern-day authorship, for example Edward Saïd's ideas of academic freedom of thought and exploration of ideas in academia. Here, we find a development from the idea of merely (though there is no 'merely' about it) cultivating and protection human potential. HE offers a step from formation of the human and the potential we know is there, to achieving new, original, and undiscovered potential. It can therefore be said that the learner's human potential in HE extends a bit further, to potentially unknown horizons. Employability as an end may be limited, for example, but the meaningful redesign of careers is a constructivist exercise. Technology in itself is a false god too, but the use of it for the purpose of maximising human creativity to new horizons could be meaningful. HE is about discovery and reach to an unknown. This also necessitates an exploration of values that guide towards the unknown.

A sweeping generalisation is that education today has firm humanist values, but humanism is a complex matter in itself, with a long historical development and various theoretical nuances. The focus on the 'human' has been problematised in various ways, generally accepting that the meaning of 'human' is contingent. It is therefore a deeply constructivist exercise, to determine what is human, then its potential, then the purpose of education in regards to human potential, what education is in itself, and to what unknown it may lead. There is also a seemingly increasing line of thought on the human as limitation. This gives rise to posthumanist or transhumanist theory, and related purposes of education. Radical posthumanist thinkers will echo Max More's *Letter to Mother Nature* (2009), a manifesto in which the biological limitations of being human are denounced in favour of new potential for being. It is a remarkable constructivist stance, to recalibrate an entire mode of being. More (2009) states the next steps will be developed 'cautiously, intelligently, and in pursuit of excellence', 'guided by critical and creative thinking'. In a paradoxical way, the ambition to leave human limitation behind still links to a belief in the human potential to become, develop, learn – perhaps towards more and greater things than we had previously imagined. While posthumanist theory has a central desire to decentre a focus on the human, it must also acknowledge a belief in the human to be able to do so: potential, in other words.

In a multi-university collaboration, Snaza et al. (2014) call for new approaches to education with the underlying purpose of abandoning anthropocentric perspectives. It poses a similar view on the narrow horizons of the human, while believing humans can be more – as posthumans, perhaps. Hence an interest in the term 'transhuman', to indicate it is a progression, rather than departure. What the posthuman curriculum might do is hard to envision, Snaza et al. (2014) write. Snaza (2013) coined the term 'bewildering' to describe this: 'a grappling toward a radically different sense of what "pedagogy" could mean' for an education system that does not purposefully intend to humanise (p.39). There is a bewildering aspect to GenAI, in terms of its speed, natural communication, and convincingness of information. Snaza (2013) poses it is good to accept that the means of education are not currently clear, because the purpose is: a focus on the 'why', rather than 'how'.

Andersson (2022) refocuses posthumanism by suggesting a continuum that entangles a conscious, interpretive subject with nonconscious cognition, which was also theorised by Katherine Hayles. It suggests the (human) subject as 'retained rather than dethroned' (Andersson, 2022). This could philosophically align with GenAI, as it proposes a conscious, interpreting human prompter interacting with a non-conscious information-generating machine. The issue would occur in allowing an attribution of interpretation and consciousness to the machine. Due to its natural communication technique, and convincing nature of information generated (even when there is a growing awareness of AI hallucinations and misinformation), this may be easily done so. That would lead to Postman's (1992) worrying stipulations of the technopoly. It reaffirms the importance of HE to forefront human potential as an end.

These continuities are becoming more widely recognised in educational theorising, suggesting students are entangled with other intelligences (some of which artificial), and that this requires a re-examination of fundamental aspects of education. Despite a pro-adoption of technology which pervades the UK strategies, the human potentiality of thinking, interpreting, criticality, and scepticism, come forward as retained elements of the humanist narrative in considering the future of education. It is hard to imagine what this education might constitute. It is 'almost unthinkable', but: 'The unthinkable, however, is a good place to begin if we might want to imagine education for a different sort of human from the one we have been thinking with for far too long' (Facer, 2023). The end of education remains human potential, and HE especially, remains the place to boldly explore new potential. There may be a cautionary note needed to the idea of unbridled pursuit of human potential, however. The ethics, critical reflection, and debate that lie at the idea of a university serve as necessary safeguards. There is also the likelihood that potential will (need to) be directed, as many educationalists have theorised more specific aims for the development of the human potential according to their interests, time, and place. Still, it works as a foundation for the unifying narrative which Postman (1996) proposed as essential for education.

6. Conclusion

This paper set out with the ambition to explore whether there might be an end of education that emerges as a common theme for HE in response to the rapid emergence of GenAI. A dominant national view leans towards technology acceptance, meaning UK HE and UK government have adopted the position of *using* the new technology, in a pro-growth and pro-innovation manner, rather than avoiding it. The transformative impact of GenAI is widely accepted as certain and inevitable, which is a determinist stance. However, the exact nature of changes still carries uncertainty, which opens a space for constructivist thinking. Much still depends on human use and decision-making.

The constructivist stance is reflected in UK HE policy-making and position statements. Most of these efforts have focused on 'how' to integrate AI in education, rather than 'why'. The latter is more important to Postman (1996). Without a purpose, education is steerless. But Postman (1996) also warns against the 'false gods' of employment, consumerism, and technology. Taking these as the purpose of education would also spell education's demise. Yet various policies and position statements show that HE may veer towards exactly those ends. This would lead to Postman's (1992) feared 'technopoly', a society where people's worldviews are exclusively formed by technology's determinist influence.

A central value which emerges, however, is human potential. Negative impact of GenAI is perceived as an interference with the human potential for critical thinking, intelligence, creativity, integrity, etc. Yet depending on use, it can also enhance the development of that potential in education. The focus on human potential as a purpose for education aligns well with the history of educational theory. However education is organised, it centres on the belief in the ability of people to become. What they (should) become varies across educationalists. Paradoxically, even in post/transhumanist theories of education, there is a belief in human potential still. So, the end of education should remain human potential (retained rather than dethroned).

Without believing in human potential, educating people becomes meaningless indeed. In HE in particular, the human potential is characterised as driving further, to new creativity, and the distant horizons that these learners (and their societies) might achieve.

Human potential can viably be an end for education, to which technologies are subjected as tools when thinking about 'how' to educate for potential. More work is needed to further articulate this 'why', before tackling 'how' to integrate GenAI in teaching, learning or assessment.

References

Andersson, I. (2022) The Subject in Posthumanist Theory: Retained Rather than Dethroned. *Educational Philosophy and Theory*, Vol 54, No 4, pp395-403.

Braidotti, R. (2013) *The posthuman*. Polity.

Brown, R. (2023) The AI generation will need more explicit preparation for graduate employment. WONKHE, November 8th. https://wonkhe.com/blogs/the-ai-generation-will-need-more-explicit-preparation-for-graduate-employment/

Crawford, A. (2024) AI's Rise Is Unstoppable Whatever Government Leaders Do. *Bloomberg UK*, March 18th. https://www.bloomberg.com/news/newsletters/2024-03-18/ai-s-rise-is-unstoppable-no-matter-what-goverment-leaders-say

Davies, M. (2024) A Lost Decade? The UK's Industrial Approach to AI. https://ainowinstitute.org/publication/a-lost-decade-the-uks-industrial-approach-to-ai#h-a-lost-decade-recurring-themes-in-uk-ai-policy

DfE (2019) Realising the potential of technology in education: A strategy for education providers and the technology industry. https://assets.publishing.service.gov.uk/media/5ca360bee5274a77d479facc/DfE-Education_Technology_Strategy.pdf

DfE (2023) Generative artificial intelligence (AI) in education. https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education/generative-artificial-intelligence-ai-in-education

DfE (2024) GenAI in education: Educator and expert views. https://assets.publishing.service.gov.uk/media/65b8cd41b5cb6e000d8bb74e/DfE_GenAI_in_education_-_Educator_and_expert_views_report.pdf

Facer, K. (2023) Rethinking the 'human' at the heart of humanist education, UNESCO, 7th July 2021/2023. https://www.unesco.org/en/articles/rethinking-human-heart-humanist-education

Felix, J. & L. Webb (2024) Use of artificial intelligence in education delivery and assessment, January 24th. https://post.parliament.uk/research-briefings/post-pn-0712/

Fleck, J. & J. Howells (2001) Technology, the Technology Complex and the Paradox of Technological Determinism. *Technology Analysis & Strategic Management*, Vol 13, No 4, pp 523-531.

Gov UK (2021) The National AI Strategy. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf

Gov UK (2022) National Data Strategy. https://www.gov.uk/guidance/national-data-strategy

Gov UK (2023) AI regulation: a pro-innovation approach. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

Hallström, J. (2020) Embodying the past, designing the future: technological determinism reconsidered in technology education, International Journal of Technology and Design Education, Vol 32, pp 17-31.

Hayles, N. K. (2017) *Unthought: The power of the cognitive nonconscious*. The University of Chicago Press.

JISC (2023) Student perceptions of Generative AI. https://repository.jisc.ac.uk/9218/1/NCAI-Students-Perceptions-of-generative-AI-Report.pdf

JISC (2024) Student perceptions of Generative AI. https://repository.jisc.ac.uk/9571/1/student-perceptions-of-generative-ai-report.pdf

King's College London (2024) King's guidance on GenAI for teaching, assessment and feedback. https://www.kcl.ac.uk/about/strategy/learning-and-teaching/ai-guidance

Kirkwood, A. (2015) Teaching and learning with technology in HE: blended and distance education needs 'joined-up thinking' rather than technological determinism. Open Learning: The Journal of Open, Distance and e-Learning, Vol 29, No 3, pp 206-221.

Lea, Gary R. (2020) Constructivism and Its Risks in Artificial Intelligence. *Prometheus*, Vol 36, No 4, pp 322–46.

McGregor, J. (2024) Understanding the Impact of GenAI On The High-Tech Industry, March 19th. https://www.forbes.com/sites/tiriasresearch/2024/03/19/understanding-the-impact-of-generative-ai-on-the-high-tech-industry/

More, M. (1999/2009) A Letter to Mother Nature: Amendments to the Human Constitution, 25th May. https://strategicphilosophy.blogspot.com/2009/05/its-about-ten-years-since-i-wrote.html

Newman, J.H. (1873/1996) *The Idea of a University*. Yale UP.

Oliver, M. (2011) Technological determinism in educational technology research: some alternative ways of thinking about the relationship between learning and technology. *Journal of Computer Assisted Learning*, Vol 27, No 5, pp 373-384.

Postman, N. (1992) *Technopoly: the surrender of culture to technology.* New York: Knopf.

Postman, N. (1996) *The End of Education*. Random House: New York.

Postman, N. (1998) Five things we need to know about technological change, Talk delivered in Denver, CO. https://www.cs.ucdavis.edu/~rogaway/classes/188/materials/postman.pdf

QAA (2023) Generative Artificial Intelligence in Education: QAA's response to DfE call for evidence, September 7th. https://www.qaa.ac.uk/news-events/news/qaa-publishes-response-to-department-for-education-generative-artificial-intelligence-consultation

QAA (2024) Navigating the complexities of the artificial intelligence era in HE. https://www.qaa.ac.uk/docs/qaa/news/quality-compass-navigating-the-complexities-of-the-artificial-intelligence-era-in-higher-education.pdf?sfvrsn=8179b281_11

Russell Group (2023) New principles on use of AI in education. https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/

Seal, M. & A. Smith (2021) *Enabling Critical Pedagogy in HE*. Critical Publishing.

Selwyn, N. (2022) The future of AI and education: Some cautionary notes. *European Journal of Education*, Vol 57, No 4, pp 531-691.

Snaza, N. (2013) Bewildering Education. *Journal of Curriculum and Pedagogy*, Vol 10, No 1, pp 38-54.

Snaza, N., et al. (2014) Toward a Posthumanist Education. *Journal of Curriculum Theorizing*, Vol 30, No 2, pp 39-55.

Tight, M. (2023) Employability: a core role of HE? *Research in Post-Compulsory Education*, Vol 28, No 4, 551-571.

UCL (2024) Using AI tools in assessment. https://www.ucl.ac.uk/teaching-learning/generative-ai-hub/using-ai-tools-assessment# AI has an integral role

University of Oxford (2024) Use of GenAI tools to support learning, January 8th. https://www.ox.ac.uk/students/academic/guidance/skills/ai-study

Saïd, E. W. (1994) Identity, Authority, and Freedom: The Potentate and the Traveler. *Boundary* 2, Vol 21, No 3, pp 1–18.

Wooldridge, A. (2023) Can Oxford and Cambridge Save Harvard From ChatGPT? *Washington Post*, August 23rd. https://www.washingtonpost.com/business/2023/08/23/can-oxford-and-cambridge-save-harvard-from-chatgpt/71bd07ae-416c-11ee-9677-53cc50eb3f77_story.html

Wyatt, S. (2008) Technological determinism is dead; long live technological determinism. *The Handbook of Science & Technology Studies* (Eds. Hackett et al.). Routledge, pp 165 - 180.

Zewe, A. (2023) Explained: GenAI. *MIT News*, November 9th. https://news.mit.edu/2023/explained-generative-ai-1109

Can Robots Laugh? An Inquiry in the Nature of AI

Eduardo Tomé[1] and Elizaveta Gromova[2]
[1]Universidade Lusófona, Lisbon, Portugal
[2]Independent Researcher, Portugal

eduardo.tome@gmail.com
lizaveta-90@yandex.ru

Abstract: In this paper we discuss the possibility of robots having a mind and being able to act like human beings and even surpass the human intelligence, and in consequence taking over the world. It is possibility that has been put forward in human history long ago, and that has been accentuated with the new advances in technology from the last few years, of which Chat GPT is the last very well-known example. We base ourselves in a literature review made on eight basic features we define as characteristic of humans, namely: Reproduction, Creation, Belonging, Citizenship, Self-Awareness, Mortality, Rationality, Humour, Feelings and Emotions. We use a plurality of databases as Google and SCOPUS. As a result, we conclude that even if robots may express themselves as humans, and may beat humans in specific activities, they lack most of the features that define human beings and most probably they will ever do. As with time and space travelling, robots that would take power on Earth are a utopia that will probably never happen, but whose pursue will be beneficial for the human race. The paper has the limitation of being only theoretical, and the originality of being based on Philosophy of Artificial Intelligence and presented in a scientific environment.

Keywords: Robots, Humanity, Soul, Inteligence~

1. Introduction

Robots and Artificial Intelligence appeared in human history first as a dream of writers, like Mary Shelley's Frankenstein (Shelley, 1818). Just fifty years before, the idea of a machine that played chess was put forward, only to be uncovered as a lie (Standage, 2002). In fact, the first time that computers managed to play chess was due to the work of Alan Turing (Copeland, 2004), the mathematical genius that had helped decoding the German Enigma military machines (Reuvers and Simon, 2010). After several decades of great development and enthusiasm, the idea of AI surpassing mankind was decisively put forward by Bolstrom, 2014. According to this very well-known and quoted author, in some decades, humans could be comprehensively surpassed by robots, generating a very big problem of control.

In this paper we want to address this question, and we start by assuming that in order to beat humans, the robots would need to be something like "improved humans". By humans we mean an entity that has autonomy and that has a mind; that autonomy enables humans to perform a vast set of tasks, many of them unexpected; to have a mind means robots should have self-awareness and the idea of life and death and of existence; ultimately, as mentioned in the title of the paper, it would mean that robots would have emotions, would cry when things go wrong and would laugh when things go better than expected. Also, humans are citizens, they do not only create but they procreate; and they live in societies, they have citizenship. We really think that at the first quarter of the 21st century, humanity is very advanced, so advanced it is pursuing big dreams, like creating robots. But in the meantime, this is done by not having in consideration the value of humans themselves as intelligent, emotional and social beings. By understanding robots (and their shortcomings) we might be able to understand and value humans better.

Having what was just said in mind, and in order to answer to the research question (What is the nature of Artificial Intelligence, is it comparable to Human Nature?), the paper will be composed by the following sections: concepts, methods and data, analysis, discussion, and conclusions.

2. Concepts

2.1 Human Nature

By Human Nature we define a being which has both mind and body. That being is capable of reproducing itself and lives in groups, called families, clubs, cities or societies, becoming citizens. Crucially, body and mind are interlinked and give humans a sense of existence and of mortality; this sense is the origin of both art and philosophy. Humans have also a deep sense of belonging, even if they may be ferociously competitive with one another. Humans are capable of rational thoughts, but they also have humour, feelings and emotions; they also have values which guide their lives. Mind and body put together make it possible to create and to procreate, meaning that humans create things that are outside their bodies and also are able to prolong the human species.

Crucially, mind and body make it possible for humans to move and to decide at their own will what to do with their own existence as well as to influence their fellow citizens. Humans perform lots of tasks every day.

2.2 Robots

Robots are objects made of metal, plastic and silicone. They do not know they exist. They cannot replicate themselves. They do not have sense of belonging. They may be extremely good at doing some specific tasks but lack capabilities to do multiple tasks. They only may move by themselves limitedly. They can create something close to art or philosophy, but because they are not aware of themselves, these works are not art or philosophy: they only look like it.

The ideas exposed in the last two sub-sections are summarized in Table 1.

Table 1: Humans versus robots

Topic	Humans	Robots
Reproduction	Natural	Extremely limited – xenobots
Belonging	Deep and necessary	Impossible
Citizenship	Decisive	Not foreseen
Self-Awareness	Essential	Impossible
Mortality	Definitive	Not foreseen
Rationality	Important	Technical Limited in scope
Humour, feelings, emotions	Decisive	Impossible
Creation	Permanent	Technical

3. Methods and Data

We used the eight ideas that compose the first column of Table 1 as keywords and we searched SCOPUS and Google for information. We obtained the data we expose in Section 4.

4. Analysis

4.1 Reproduction

The idea that robots may reproduce themselves is essential if we consider them to be comparable to the human race or any other living species. Very recently, scientists have announced the creation of xenobots (Ramanujam, Rasikannan, Anandhalakshmi, Kamal, 2022) as "A Remarkable Combination of an Artificial Intelligence-Based Biological Living Robot". These xenobots are the first "living robots" and they are made from stem cells from frogs (Young, 2020). Crucially, they may reproduce themselves (Hunt, 2021*).*

> *"The xenobots used "kinetic replication" – a process that is known to occur at the molecular level but has never been observed before at the scale of whole cells or organisms, Bongard said. With the help of artificial intelligence, the researchers then tested billions of body shapes to make the xenobots more effective at this type of replication. The supercomputer came up with a C-shape that resembled Pac-Man, the 1980s video game. They found it was able to find tiny stem cells in a petri dish, gather hundreds of them inside its mouth, and a few days later the bundle of cells became new xenobots" (Hunt, 2021).*

Scientists believe xenobots may be used as helpers in scientific experiments:

> *"this combination of molecular biology and artificial intelligence could potentially be used in a host of tasks in the body and the environment, according to the researchers. This may include things like collecting microplastics in the oceans, inspecting root systems and regenerative medicine." (Hunt, 2021).*

Xenobots are seen as a new way of describing the machine metaphor in biology (Barwich and Rodriguez, 2024), as new bioinformational way of understanding AI (Gautam and Deverakonda, 2024), as new way of addressing embodied cognition (Bonhard, 2023); finally, xenobots may be used in drug discovery (Solanki, Mahant, Patel, Koria and Patel, 2022) and in spaceflight (Kiesling, Maitland and Simmonds, 2022).

The achievement we described is remarkable. Having said that, the idea of robots being able to reproduce themselves as a species is not foreseen, and indeed was not even put forward by very well-respected science-

fiction writers (such as Asimov, 1950) for whom robots would be always built by humans. In fact, Asimov coined the Three Laws of Robotics:

> *"The First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm;. The Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law; The Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law." (Asimov, 1952).*

4.2 Belonging

As seen above in the previous sub-section, robots have not been considered as capable of reproducing themselves in order to create their own species. However, quite recently, it was announced that a woman, Spanish artist from Barcelona Alicia Framis, "married a hologram" (Velasquez and Ferrant, 2024). This was preceded by a less mediatised occasion when a Japanese man (Mr Akihiko Kondo) unofficially "married" fiction character Hatsune Miku (Japan Times, 2018). Mrs Framis' actions have been described as a "performance" (Velasquez and Ferrant, 2024), and Mr Kondo was reported to have made his decision after being rejected by a woman and falling into depression – somehow Mrs Miku saved him from depression. More recently, in 2023, Mr Kondo created the General Incorporated Association of Fictosexuality. The idea of somebody falling in love with a computer-generated mind had been also put forward in the movie "Her" (Jones, 2013). In this movie, a relationship is developed between a man and an artificially intelligent virtual assistant personified through a female voice: the man character is played by Joaquin Phoenix, and Scarlett Johannsson gives voice to the AI device. More prosaically, several personal assistants have been developed before the pandemics like Siri from Apple, Google Assistant, Alexa, and Cortana from Microsoft and Alice from Yandex; all these "assistants" are very developed software that may interact and "talk" to the user while helping him; one of the main and more useful features of these programs relate to the GPS location, enabling driving and movement in so-called smart cities. All in all, these "assistants" relate to loneliness and to what the meaning of social life is at the end of the first quarter of the 21st century. It is not by accident that in the UK, a Minister of Loneliness was appointed in 2018 (Daley, 2018). There is something that makes one remember the Eleanor Rigby song (Lennon and McCartney, 1966) in the discussion about robots as social elements: where are the lonely people, or should they have a robot not to feel alone?

4.3 Citizenship

Some authors have discussed the role of robots in governance (Hilb, 2020) defining five scenarios of artificial governance: assisted, augmented, amplified, autonomous and autopoietic intelligence. Also, the possibility of robots paying taxes (Kim, 2024; Oberson, 2024), when imported (Costinot and Werning, 2023) or when used as substitute of labour (Chowdhury, 2022) was analysed. It was found that robots are real source of income for governments in times of crisis, and these funds may be used in the benefit of the same persons that may be disadvantaged and unemployed by the use of robots themselves.

However, the idea of robots having an ID card and citizen rights has not been put forward, because no robots have been so developed as to be considered a person. One known exception to this situation is Sofia, a robot created by Hanson Robotics and that, on October 25th 2017, at the Summit on Future Investment Initiative held in Riyadh, was declared as official citizen of Saudi Arabia (Fernandes, 2017). Sophia gave interviews to major outlets like the CBS, CNBC or The Guardian, but the idea of having "citizenship" was seen as a publicity stunt by Saudi Arabia (Jaynes, 2019) and a way to attract foreign investment in AI (Nease, 2020); other concerns happened with vote, marriage or murder (Maza, 2017). The citizenship award led to a "legal quandry" (British Council, 2020), and in 2017 even the European Parliament considered "granting of 'electronic personhood' to the most advanced machines to ensure their rights and responsibilities" (British Council, 2020) Nevertheless, those plans *"were considered as 'ideological, nonsensical and non-pragmatic by experts in medicine, robotics, AI and ethics*" (British Council, 2020).

4.4 Self-Awareness

The possibility of robots that have sense of existence has been portrayed in the movies long ago: Robocop was a humanoid (Verhoeven, 1987), and Number 5 was a robot who feared to be disassembled more than everything (Badham, 1985). However, these brilliant ideas were just good and entertaining science fiction and art. In scientific terms, the question is much more nuanced.

In the first semester of 2023-4, the Oxford Philosophical Society produced an online course on the "Philosophy of Artificial Intelligence". Among the topics analysed in those sessions were the following: "Can a robot be a

philosopher?" by Peter Gibson on November 1, 2023 (Gibson, 2023a) and "Could a robot be an artist?" by Peter Gibson on December 20, 2023 (Gibson, 2023b). In both lectures it was underlined that robots lack an essential feature that artists and philosophers must have, which is self-awareness. Indeed, both Art and Philosophy are major constructs of an intellectual entity (the mind) in relation with its environment, and they originate from the reflection or reaction of the self-conscious mind in relation to that environment. From this idea, two very interesting conclusions are derived: first, even if robots may produce something that formally may be called art, such as novels, painting, music and even ballet (Gibson, 2023b), all this is made in a technical, mechanical and non-dramatic manner, because we are only talking about algorithms that have been programmed and built by humans, that may be switched of by humans, and that therefore do not exist as personas, with identity and soul outside that human creation. Therefore, ChatGPT or any other very performant AI algorithm could be a very good assistant to a given philosopher, but it would never be considered as a philosopher even if the machine could write philosophical essays, because all the construction would be only technically based, and it would lack the sense of ontology that is need in all philosophies – even if the robots can define some possible paths to humanity, it is impossible for them to understand the "peripatetic mantra" – to philosophe is to be in the way – because they are not themselves aware of the existence of the way. Therefore, no awareness means no drama as artists and no ontology as philosophers, and therefore even though both those types of work made by algorithms look "like art" and "like philosophy", they should only be considered as second rate in comparison to those made by humans.

4.5 Mortality

Because robots are not self-aware of their own existence, they do not feel to be mortal. Even with the recent advances of biotechnology, like biobots, the dilemmas of life and death still do not apply to them (Levin, 2021). This idea is in itself quite important because it limits the possibilities of robots as entities. Also, and quite interestingly, it has been defended that if robots could be mortal (Delbert, 2019), they would be more performant, and more able to make complex decisions, because the "survival mode" is a way to improve performance (Man, Damasio, 2019).

4.6 Rationality

The possibility of robots making very complex calculations has been highlighted by chess algorithms like Alpha Zero (Knapton and Watson, 2017) and also by scientific software like SPSS, Stata, e-views or InVivo. Quite crucially, all those very strong programs outperform the human mind even in chess. But also quite importantly, not only have they been programmed by humans, so far they have not been able to out-program themselves consciously. Therefore, even if there is a possibility that a robot could generate harm, by malfunctioning or by bad coding, the idea that robots could consciously create harm in the world (killings, murders, etc.) has been, so far, considered only by science fiction, generating Asimov's laws (Asimov, 1952), and Bostrom's warnings on control (Bostrom, 2014).

There is also an important caveat that must be exposed in relation with this topic of rationality, and it is the problems of complexity and uncertainty. This means that life is more diverse than chess and mathematical problem solving, and humans perform a very vast level of activities every day; at the current stage of development, it would take an enormous number of robots to compete with the daily performance of the human brain. Also, life is not as predefined as chess and maths. On the contrary: life is full of uncertainty in all aspects – uncertainty of what to do, when to do, and how to do – and quite surprisingly the human brain can adjust to all these types of uncertainties and manage them quite well, and almost instantly in most cases. It will take a big advance in robots to be able to deal with both the complexity and the uncertainty of human life.

4.7 Emotions, Humour and Feelings

In 2019, three researchers published a paper (Rombão, Tomé and Ribeiro, 2019), in which they made a survey about the nature or robots and their social importance, and they found that on the one hand, ordinary people believe that robots may be very useful as tools in organizations, but on the other hand, they consider robots lack one decisive aspect of humans – they do not have emotions, or feelings.

Quite recently, Sahota placed the question of feelings by robots in the realm of imagination, and so did Kenyon (2024).

So for the time being robots are emotionless – and, answering the question in the title of the paper, they cannot laugh. This fact decisively limits what robots are or may be. They can create works that seem to be art, they may solve problems, but they are just tools of metal, plastic and silicone. As we know body and soul are essential for

human development, this means robots will never be able to compare with humans and humans will have superiority over them.

4.8 Creation

As mentioned in subsection 4.4, the idea that AI can create something like books or paintings has been advocated. Vlaad, 2024 considers that the notion of "author" should be clarified. Also, according to the same author, the images developed by AI are in fact conjunct works of the human developer that created the algorithm and the algorithm which needed the human instructions to create the piece of art; as a consequence of that possibility, separate competitions for AI-generated art works are envisaged.

5. Discussion

For the economic world, the importance of robots has been put as almost absolute. As previously noted in subsection 4.3, artificial governance has been considered, just like the effect of robots in taxes. Moreover, it has been almost consensual that robots may be used in organizations and that they will change the nature of work (Selenko, Bankins, Dhoos, Warbuton and Restubog 2022). An interesting caveat in this context has been to go from a position of "men and against the machine" to a position of "man with machine against man without machine". The question of power in the organizations has also been addressed; in an international conference held just before the pandemic in July 2019, the first author of this paper heard an engineer defining organizations as "a set of servers of carbon and silicone connected in real time"; this was, however, before COVID-19 and ChatGPT questioned the existence of both presential classes and lecturers.

Moreover, the idea that robots can become dominant in the world may hide some sort of political agenda, meaning the image of a world in which very few will dominate all the others with the support of robots. This also may be driven by an idea of exploitation of the labour force and of some hard version of Social Darwinism, in which the robots would be the ultimate weapon of the fittest in order to assure their own survival.

Finally, it has been advocated that IT developers are the persons that mostly advocate the agenda of robots' dominance in the world, because they would be somehow dominant in the world if this happens – a bit like accountants defending accountability as the second more important activity of organizations because without correct accounts nothing will work – so, some truth originating a big fable.

The behavioural side of robots is still very small, and they are still in the hands of programmers who are not able to develop the machines in order to have the eight characteristics we defined above as "human".

6. Conclusion

The quest for artificial intelligence is parallel to the idea of travelling through space and time. For the time being, and most probably as a whole, Humanity will not be able to create robots with brains and will not be able to surpass the speed of light and arrive at other worlds. Having said that, even if the ultimate goals might not be achieved, in the meantime, and during the quest, much good, wealth and health may be and has been generated. Therefore, as a whole, and in conclusion, even if robots cannot surpass humanity, they may be very beneficial for humans, in particular if they can be used wisely. As with all the machines and devices, it is not the tool itself that is the fundamental question: it is how we use that.

This paper is essentially reflexive, a bit speculative, and written like the old-fashioned essays with no use of technological tools of analysis, with the exception of the word processor or the search engine. It was also proofread by a human rather than a software. However, the questions we addressed are worth being discussed, because they touch deep in the nature of humans – so, by understanding the limits of robots, as the limits of time and space, we can manage to understand humanity and its strengths and limitations.

Moreover, we feel that philosophy and reflective thoughts are essential in scientific conferences, particularly in a time where technology creates so many possibilities, as well as so many ways of happiness and prosperity.

References

Asimov (1950) I Robot, DoubleDay
Asimov (1952) Turnaround Double Day.
Badham J. (1985) Short Circuit. Movie.
Barwich, A.-S. Rodriguez, M.J. Rage against the what? The machine metaphor in biology Biology and Philosophy , 39(4), 14 2024

Bongard, J. From rigid to soft to biological robots: How new materials are driving advances in the study of the embodied cognition Artificial Life and Robotics , 28(2), pp. 282–286
Bostrom N. (2014) Superintelligence: Paths, Dangers, Strategies, ISBN 978-0-19-967811-2
British Council (2020). "Should robots be citizens?". BritishCouncil.org. Archived from the original on January 9, 2020. Retrieved March 7, 2023.
Chowdhury S. (2022) Should machines be taxed like people? LSE Business Review https://blogs.lse.ac.uk/businessreview/2022/11/24/should-machines-be-taxed-like-people/
Copeland, B. Jack (2004). The Essential Turing. Oxford University Press ISBN 978-0-19-825079-1.
Costinot, A. Werning, I (2023) Robots, Trade, and Luddism: A Sufficient Statistic Approach to Optimal Technology Regulation, The Review of Economic Studies, Volume 90, Issue 5, October 2023, Pages 2261–2291, https://doi.org/10.1093/restud/rdac076
Delay J. (2018) The U.K. Now Has a "Minister for Loneliness." Here's Why It Matters https://www.smithsonianmag.com/smart-news/minister-loneliness-appointed-united-kingdom-180967883/
Delbert C (2019) Robots Need to Know They Can Die at Any Minute, Just Like the Rest of Us. Popular Mechanics https://www.popularmechanics.com/technology/robots/a29831507/robots-mortality/
Fernandes, J.V. Robots CitIzenship and gender (in)equality: the case of Sophia the Robot in Saudi Arabia Janus.net , 2022, pp. 51–64
Gautam, R.D. Devarakonda, B. (2024) Towards a bioinformational understanding of AI AI and Society , 39(2), pp. 491–513 2024
Gibson P. (202a) Could a robot be a philosopher? Oxford Philosophical Society
Gibson P. (2023b) Could a Robot be an Artist? Oxford Philosophical Society
Hilb, M. Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance. J Manag Gov 24, 851–870 (2020). https://doi.org/10.1007/s10997-020-
09519-9
Hunt, K. (2021) World's first living robots can now reproduce, scientists say, CNN https://edition.cnn.com/2021/11/29/americas/xenobots-self-replicating-robots-scn/
Japan Times (2018) Japanese man 'marries' virtual reality singer Hatsune Miku hologram http://https//www.japantimes.co.jp/news/2018/11/12/national/japanese-man-marries-virtual-reality-singer-hatsune-miku-hologram
Jaynes, T. L. (2021). "Citizenship as the exception to the rule: an addendum". AI & Society. 36 (3): 911–930. doi:10.1007/s00146-020-01105-9. PMC 7665966. PMID 33223621.
Jones S (2013) Her, movie.
Kenyon M. (2024) An AI walks into a bar... Can artificial intelligence be genuinely funny? https://www.bbc.com/future/article/20240724-can-artificial-intelligence-be-genuinely-funny
Kiesling, D. Maitland, P. Simmons, K.L. (2022) Xenobots Applications for Spaceflight Proceedings of the International Astronautical Congress, IA , 2022-September
Kim, Y.-S. (2024) Is it possible to impose a tax on Artificial-Intelligence robots? Eurasian Legal Systems in a World in Transition: Economic prosperity or disparity, and the return of politics in international law, pp. 325–335
Knapton S. Watson L (2017) Entire human chess knowledge learned and surpassed by DeepMind's AlphaZero in four hours https://www.telegraph.co.uk/science/2017/12/06/entire-human-chess-knowledge-learned-surpassed-deepminds-alphazero/
Lennon J. Mc Cartney P (1966) Eleanor Rigby, song.
Levin M. (2021) Life, death, and self: Fundamental questions of primitive cognition viewed through the lens of body plasticity and synthetic organisms Biochemical and Biophysical Research Communications , 564, pp. 114–133 2021
Man, K., Damasio, A. Homeostasis and soft robotics in the design of feeling machines. Nat Mach Intell 1, 446–452 (2019). https://doi.org/10.1038/s42256-019-0103-7
Maza, Cristina (October 26, 2017). "Saudi Arabia gives citizenship to a non-Muslim, English-Speaking robot". Newsweek. Archived from the original on November 8, 2017. Retrieved October 26, 2017.
Nease, S (2020). "Citizen Sophia: Robots and the Future of Saudi Arabia". Penn Political Review. University of Pennsylvania. Retrieved March 7, 2023.
Oberson, X.(2024) Taxing artificial intelligence Taxing Artificial Intelligence, pp. 1–250
Ramanujam, E. Rasikannan, L. Anandhalakshmi, P.A. Kamal, N.A (2022(Xenobots: A Remarkable Combination of an Artificial Intelligence-Based Biological Living Robot International Journal of Sociotechnology and Knowledge Developmen , 14(1), pp. 1–11
Reuvers, Paul; Simons, Marc (2010), Enigma Cipher Machine, retrieved 22 July 2010
Rombão, M. ., Tomé, E., & Ribeiro , C. . (2020). Will robots have the capacity to replace Mankind? Survey from Portugal. *Gestão E Sociedade*, *14*(37), 3417–3438. https://doi.org/10.21171/ges.v14i37.3084
Sahota N (2024) AI And Emotion: Imagining A World Where Machines Feel. Forbes https://www.forbes.com/sites/neilsahota/2024/01/22/ai-and-emotion-imagining-a-world-where-machines-feel/
Selenko, E., Bankins, S., Shoss, M., Warburton, J., & Restubog, S. L. D. (2022). Artificial Intelligence and the Future of Work: A Functional-Identity Perspective. Current Directions in Psychological Science, 31(3), 272-279. https://doi.org/10.1177/09637214221091823
Shelley M (1818) Frankenstein; or, The Modern Prometheus

Solanki, N. Mahant, S. Patel, S. Koria, H. Patel, A. (2022) Xenobots: Applications in Drug Discovery Current Pharmaceutical Biotechnology , 23(14), pp. 1691–1703 2022

Standage T. (2002), The Turk: The Life and Times of the Famous Eighteenth-Century Chess-Playing Machine (New York: Walker), 22–23.

Velázquez J. Farrant T. ()2024) Meet Spanish artist Alicia Framis: The first woman to marry a hologram https://www.euronews.com/culture/2024/01/03/meet-spanish-artist-alicia-framis-the-first-woman-to-marry-a-hologram

Verhoeven P. (1987) Robocop movie.

Vlaad, S. (2024) A portrait of the artist as a young algorithm Ethics and Information Technology, 26(3), 58

Young J. (2020) Meet the xenobot: world's first living, self-healing robots created from frog stem cells CNN https://edition.cnn.com/2020/01/13/us/living-robot-stem-cells-intl-hnk-scli-scn/

An Analysis of Knowledge Management Changes Through Artificial Intelligence with Probst's Model

Burak Toptas
Nuremberg University of Applied Sciences, Nuremberg, Germany

Toptasbu84460@th-nuernberg.de

Abstract: In recent years, the integration of Artificial Intelligence (AI) into Knowledge Management (KM) has led to transformational changes. These changes have significantly enhanced traditional KM processes. To identify how AI technologies improve and reshape knowledge processes, this study conducted a systematic literature review. The review identified AI technologies suitable for each of Probst's building blocks, which outline the eight central KM processes. The research reveals a wide range of AI technologies, including machine learning, natural language processing and chatbots such as ChatGPT. These technologies can be applied in different domains and introduce innovative approaches to improve KM processes. Based on the AI technologies analysed, this study proposes a four-stage model to support the documentation and application of best practices and lessons learned. The model is designed to enhance the knowledge development process and aims to document and secure key project developments in the long term. A further objective was to analyse which KM process is most affected by chatbots. The findings indicate that chatbots have the potential to transform the use of knowledge in organisations. They act as facilitators by breaking down existing barriers, foster an open culture of knowledge sharing, streamline workflows and increase the accessibility of knowledge. The study also examines the broader changes that AI will bring to KM and forecasts the sixth generation of KM. It draws on Bencsik's (2021) evolutionary and revolutionary perspectives that specifically forecast this next generation. The study shows that AI not only enhances existing KM processes but also has the potential to fundamentally disrupt traditional methods and approaches. These findings underline the need for future research to explore the effective integration and scalability of AI technologies in real-world KM environments. This will help ensure that their long term impact and potential benefits are fully understood across different industries and organisational contexts.

Keywords: Knowledge management, Artificial intelligence, Probst's Building Block Model, Innovation, AI technologies for knowledge management

1. Introduction

The origins of knowledge management (KM) can be traced back to Michael Polanyi's (1966) work, “The Tacit Dimension”. Initially, Polanyi's work had no significant impact on the discipline, but it has since become a cornerstone of KM. The foundations of artificial intelligence (AI) were laid earlier by Alan Turing in 1950 through his article “Computing Machinery and Intelligence”, which introduced the “Turing Test” to assess machine intelligence. This test shaped the core principles and objectives of AI, profoundly influencing the field to this day. These historical milestones laid the groundwork for extensive literature on KM and AI. Their development has fostered a strong symbiosis between the two fields, which are now more interconnected than ever. The primary goals of KM are to capture, utilise and develop knowledge. Effective KM provides organisations with a sustainable competitive advantage (Davenport, 1998). This can lead to higher quality, increased creativity and improved efficiency. Knowledge is divided into explicit (documented and codified) and tacit (based on personal experiences and hard to articulate) forms (Nonaka and Takeuchi, 1995). While knowledge management systems (KMS) generally focus on explicit knowledge, effectively transforming tacit knowledge into a shareable format is a significant challenge that limits their overall effectiveness. The key challenge is to capture and structure tacit knowledge to make it shareable (Nonaka and Takeuchi, 1995).

In today's dynamic business environment, the development and implementation of effective KMS are crucial to ensure adaptability and competitiveness (Bencsik, 2021). The integration of AI with KM has recently emerged as a key driver of transformation, which enables machines to learn, apply and develop knowledge to improve accessibility for employees. AI and KM complement each other: KM aids to understand and structure knowledge, while AI enhances efficiency in the expansion, use and development of knowledge (Vajpayee and Ramachandran, 2019). This intersection enables organisations to drive innovation and streamline processes by leveraging the decision-making and automation capabilities of AI.

Building on these foundations, existing research has explored the role of AI in specific KM processes. However, it does not comprehensively cover the full scope of KM processes. Mohammad et al. (2022) focus on the relationship between AI technologies and KM processes. They identify AI technologies applicable to six KM processes and establish a significant link between knowledge development and data mining. Their study underscores the challenge of fully integrating AI into KM. Bencsik (2021) extends this perspective through an analysis of the KM processes represented in Probst's building block model (Probst, Raub and Romhardt, 2012).

The focus is on the prediction of the success of innovation and the assurance of strategic feasibility through informed management decisions. The study provides valuable insights into the potential applications of AI in different KM processes. Furthermore, Bencsik (2021) lays the groundwork for broader changes in KM through AI and proposes the future role of AI in the sixth generation of KM. Bencsik (2021) highlights a research gap in this area, which emphasises the need for detailed exploration of KM's future possibilities.

Bencsik (2021) categorises the evolution of KM into five generations. Each generation reflects shifts in focus and methodology. The *first generation* centred on the use of IT for knowledge creation and management, though it proved inadequate for managing tacit knowledge. The *second generation* recognised diverse types of knowledge, including knowledge-based, experience-based and problem-solving approaches. The challenge was the codification of tacit knowledge for wider distribution. The *third generation* viewed KM as a network to mobilise knowledge and enhance performance through maintenance and networking. The *fourth generation* positioned knowledge as a capital factor, aims to quantify and utilise knowledge resources effectively. The *fifth generation* explored the interplay between corporate competitiveness and innovation, focusing on the valuation and quantification of human capital.

These generations represent the natural evolution driven by expert thinking, organisational practices and technological advances. Building on this framework, Bencsik (2021) proposes two potential future directions for the sixth generation of KM. The first is an *evolutionary* approach, which returns to the initial stages of KM but at a higher level, where IT regains importance and AI overshadows previous focuses. This scenario mirrors the first generation but incorporates advanced IT capabilities like AI to drive KM evolution on an expanded technological scale. The second approach is *revolutionary* in its aim to fully utilise the capabilities of AI to bring the entire KM process to a new level of organisational efficiency. In this approach, AI-based solutions are integrated into the KM processes, to enable faster, more efficient and successful interventions or tasks.

This structured overview not only highlights the historical evolution of KM, but is also essential for the anticipation of the characteristics and innovations of the next, sixth generation. It shapes our understanding of how KM may evolve with advances in AI and organisational needs. Therefore, this study investigates changes in KM through AI based on Probst's building block model, which outlines the entire KM process (Probst, Raub and Romhardt, 2012). As outlined by Brocke et al. (2009), a systematic literature review was conducted to identify new AI technologies that improve KM processes. This research includes four key contributions. First, it identifies and examines AI technologies suitable for each KM process, presented in a comprehensive tabular format. Second, it proposes a four-stage model to support the creation of best practices and lessons learned for the knowledge development process. Third, it provides a predictive analysis of the anticipated sixth generation of KM. Fourth, it analyses the KM building block most affected by chatbots like ChatGPT.

2. Background: Probst's Building Block Model

The Probst building block model is essential for understanding KM processes, as illustrated in Figure 1. This model comprises eight blocks in a cyclical sequence that links the strategic and operational aspects of KM (Probst, Raub and Romhardt, 2012). The outer cycle includes strategic blocks essential to secure the organisation's long-term core competencies, while the inner cycle includes operational blocks that drive these strategic objectives. Each block addresses specific problem areas that may disrupt knowledge flow. The building blocks are described as follows according to Probst, Raub and Romhardt, 2012.

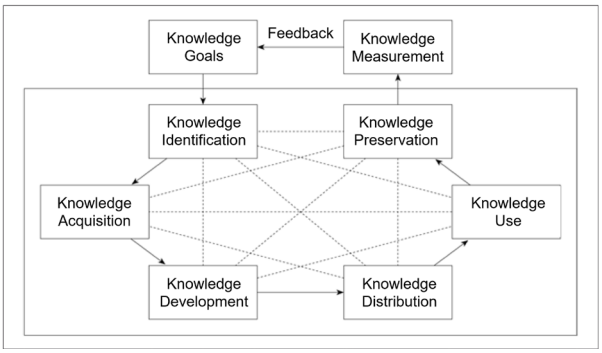


Figure 1: Probst's building block model (Source: Based on Probst, Raub and Romhardt, 2012, p. 34)

Defining *knowledge goals* provides the direction for KM activities across three organisational levels. Normative goals foster a knowledge-conscious culture. Strategic goals define core organisational knowledge and future

competencies. Operational goals ensure KM implementation at every organisational level. Effective knowledge identification is achieved by building structures that maintain transparency, reduce inefficiencies, prevent redundancies and aid employees in information retrieval. Knowledge acquisition adds value through external sources, such as hiring specialists, acquiring companies, or purchasing software and repositories to enhance internal resources. Knowledge development aims to create new skills, ideas and processes that foster innovation and employee creativity. Knowledge distribution emphasises that effective transfer extends beyond the mere exchange of information. It targets organisational knowledge to specific employee groups and between individuals, teams or workgroups. The value of KM becomes tangible in knowledge use, as knowledge only generates value when actively applied. Employees accept and apply knowledge based on evident benefits. Knowledge preservation maintains accessible and relevant skills through systematic storage and updates. Finally, knowledge measurement evaluates the effectiveness of KM initiatives, ensures that knowledge objectives are aligned with organisational outcomes and that KM activities are implemented effectively.

3. Methodology

This study follows the systematic literature review approach outlined by Brocke et al. (2009). Therefore, keyword lists for “Knowledge Management” and “Artificial Intelligence” were developed, with additional terms derived from foundational literature to refine the search. The keyword lists are illustrated in Table 1.

Table 1: Keyword lists for literature searches (Source: Table created by author)

Keyword list 1	"Knowledge Management" AND ("Artificial Intelligence" OR "Artificial Neural Network*" OR "Deep Learning" OR "Machine Learning" OR "Natural Language Processing")
Keyword list 2	"Knowledge Management" AND ("Chatbot*" OR "ChatGPT" OR "Large Language Model*")
The asterisk (*) in keyword list 2 acts as a wildcard to capture both singular and plural forms of key terms.	

The literature search was conducted in eight databases: *ACM Digital Library, IATED Digital Library, PubMed, Web of Science, ERIC, SpringerLink, ScienceDirect* and *IEEE Xplore*. This was supplemented by forward and backward searches to include related work (Brocke et al., 2009). The inclusion criteria for this study encompasses peer-reviewed publications from journals, conference proceedings and books published in the last five years. These publications must specifically address the intersection of KM and AI applications. Initially, only these inclusion criteria were applied and studies that did not meet these criteria were excluded. The search was conducted from October to November 2023 and initially yielded 745 hits. This number was reduced to 670 after the removal of duplicates. Based on the inclusion criteria, 69 articles were identified. The forward and backward search identified a further 6 articles. A total of 75 articles were identified for in-depth analysis. As a result, 15 articles were selected for inclusion in this study. Figure 2 illustrates the multi-stage systematic literature review process adopted for this study.

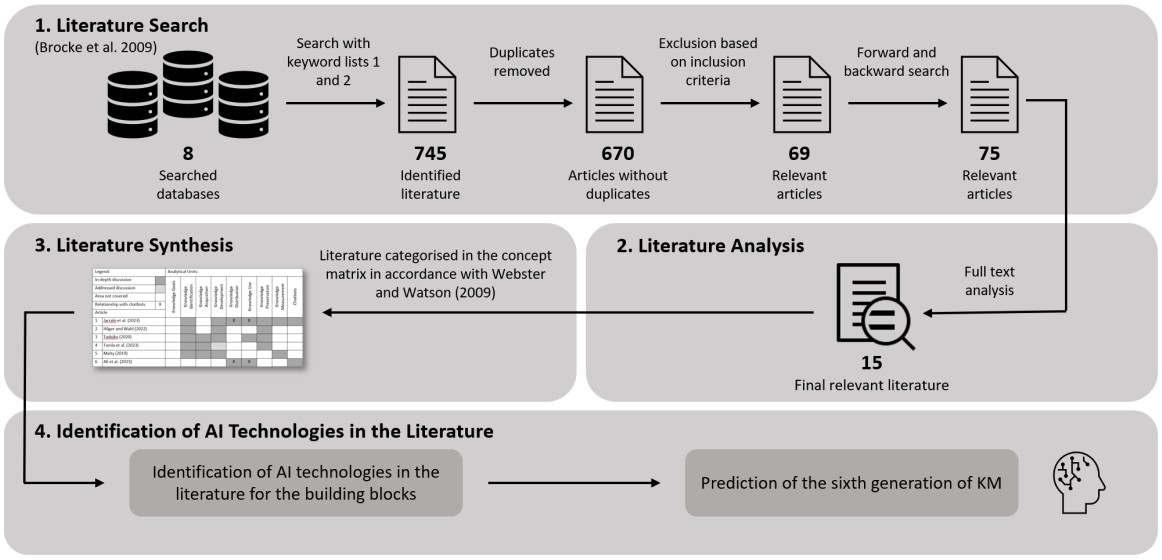


Figure 2: Overview of the literature review (Source: Figure created by author)

4. Results

This section presents the results of the literature review. The primary focus is on the analysis and synthesis of the literature.

4.1 Literature Analysis and Synthesis

The synthesis of the literature follows the methodology proposed by Webster and Watson (2002). It focuses on the identification of thematic patterns and theoretical perspectives to understand the interaction between KM and AI. The concept matrix in Table 2 is segmented in accordance with Probst's building block model. It uses colour coding to indicate the depth of discussion on AI technologies within KM processes. The matrix also positions chatbots alongside KM building blocks to emphasise the direct relationship with these technologies. An additional marker, represented by an "X", illustrates this link. For example, an "X" for knowledge development indicates that chatbots are discussed in the literature as supportive AI technologies for knowledge development. These colour differentiations are described in the legend.

Table 2: Concept matrix of the analysed literature (Source: Table created by author based on Webster and Watson, 2002, p. 17)

Legend:		Analytical Units:								
In-depth discussion		Knowledge Goals	Knowledge Identification	Knowledge Acquisition	Knowledge Development	Knowledge Distribution	Knowledge Use	Knowledge Preservation	Knowledge Measurement	Chatbots
Addressed discussion										
Area not covered										
Relationship with chatbots	X									
Article										
1	Jarrahi et al. (2023)					X	X			
2	Hilger and Wahl (2022)									
3	Tadejko (2020)									
4	Tamla et al. (2023)									
5	Maity (2019)									
6	Ali et al. (2023)					X	X			
7	Abubakar et al. (2019)									
8	Korzynski et al. (2023)			X	X	X	X			
9	Owoc, Sawicka and Weichbroth (2021)								X	
10	Bilgram and Laarmann (2023)				X		X			
11	Majumder and Dey (2022)									
12	Masood and Hashmi (2019)									
13	Botega and da Silva (2020)									
14	Bencsik (2021)									
15	Paschen, Kietzmann and Kietzmann (2019)									

The analysis in Table 2 demonstrates a strong emphasis on certain KM processes, such as *knowledge identification* and *development*, where AI technologies play a significant role. This shows that AI technologies for these processes are widely recognised and discussed. The consistent presence of chatbots across five KM processes (marked with an "X") underscores their growing importance and potential to improve KM processes. However, it is evident from Tables 2 that no AI technologies are in-depth discussed in the literature related to *knowledge goals*. This absence is noteworthy. Probst, Raub and Romhardt (2012) emphasise that these goals need to work synergistically across all corporate levels. This synergy is crucial to effectively contribute to corporate goals. It remains a challenge to integrate AI in the formulation of knowledge goals. The current capabilities of AI are exceeded by the complex nature of precise and specific knowledge goal formulation. This indicates potential areas for future research.

4.2 AI Technologies Within the KM Processes

To provide a clearer overview of the interaction between AI technologies and KM processes, Table 3 categorises these technologies identified in the reviewed literature. This table adopts a simplified presentation style where each KM process is analysed for its connection with specific AI technologies. Instead of overburdening the table with direct citations, article numbers are used that correspond to those in the concept matrix in Table 2.

Table 3: Presentation of AI technologies across all building blocks (Source: Table created by author)

AI Technologies	Knowledge Goals	Knowledge Identification	Knowledge Acquisition	Knowledge Development	Knowledge Distribution	Knowledge Use	Knowledge Preservation	Knowledge Measurement
AI-based KMS			5	13				
AI-based virtual trainers				1, 5, 9				5
Artificial neural networks (ANNs)		7						7, 14
Data mining				11			9	
Deep learning							1	
Expert systems							14	
Image processing		3		3			3	
Intelligent agents		5	5					
Intelligent search with NLP					12			
Chatbots			8	8, 10	1, 6, 8	1, 6, 8, 10		9
Machine Learning (ML)		3	3	3				3
Named entity recognition (NER)		4					4	
Natural language processing (NLP)		2, 12	3, 4, 14, 15	2, 3, 15		3	2, 3, 9, 12	
Optical character recognition (OCR)		12			12		12	

Note: The numbers in the cells refer to article indices that discuss the AI technologies in the context of the KM process. These indices correspond to the detailed list provided in the concept matrix in Table 2.

The analysis of Table 3 highlights the diverse use of AI technologies across KM processes. Notably, NLP and chatbots are predominantly used in five different KM processes. This underlines a significant reliance on these technologies to improve the efficiency of KM.

4.3 Application of AI Technologies in KM

After the identification of AI technologies for seven KM process in Table 3, this section presents selected applications of these technologies. This approach provides a foundation for understanding the potential impact of AI in KM and supports the forecast of the next generation of KM.

4.3.1 Search enhancement with NLP

NLP technologies enable systems to understand and analyse the meaning and context of text inputs from users. The integration of NLP with knowledge graphs (KG) enhances the capabilities of intelligent search and retrieval functions, which leads to more powerful search functions (Hilger and Wahl, 2022). These advanced NLP search capabilities address a common problem: employees are often faced with overwhelming results when searching relational databases such as NoSQL with a single keyword, which requires extensive filtering. The integration of NLP-based intelligent search into existing KMS can significantly improve the user experience by allowing search queries to be formulated in natural language (Masood and Hashimi, 2019). It enables employees to navigate knowledge repositories and extract relevant results from both explicit and implicit contexts.

In the context of *knowledge identification*, we highlight the importance of text analysis tools based on NLP. These tools are useful for the identification and reinforcement of weak points in organisations and the discovery of unknown patterns in large amounts of information (Hilger and Wahl, 2022; Masood and Hashimi, 2019). For organisations that aim to fully identify their knowledge resources, these technologies are essential. Named entity recognition (NER) greatly enhances the effectiveness of text analysis tools. NER identifies and contextualises specific entities in text, making overlooked information accessible (Hilger and Wahl, 2022).

4.3.2 Transformation of unstructured data into searchable information

Organisations often face the challenge of managing large volumes of unstructured data such as scanned documents, PDFs or images that are not searchable. Optical character recognition (OCR) is recommended to solve this problem. This technology extracts text from scanned data and stores it in a searchable digital form (Masood and Hashimi, 2019). After the extraction process, the unstructured data can be transformed into a searchable document by using the text to create a search index. Tools such as Azure Search Index can then be used to index and structure the content of this data, in preparation for more efficient search queries. In addition to OCR, image processing is a crucial technology for structuring unstructured data such as images. These AI technologies enable effective organisation of images through tagging and categorising, which facilitates long-term storage and retrieval of images (Tadejko, 2020). Important information is extracted from images and made usable through features such as object, scene, face and text recognition. This ensures that visual knowledge is not only stored but also accessible for future use and decision-making processes.

4.3.3 Cloud-Based AI technologies

Cloud-based AI services from providers such as Amazon AWS, Microsoft Azure, Google Cloud and IBM Watson offer advanced ML and NLP tools (Tadejko, 2020; Tamla et al., 2023). For instance, Amazon Comprehend uses NLP to analyse document content without pre-processing. This makes it valuable for assessing large textual datasets such as social media responses and customer service transcripts through sentiment analysis (Tadejko, 2020). IBM Watson NLU is designed to classify data, extract insights, perform syntax analysis and discover relationships between entities. These capabilities support the recognition and interpretation of clues in textual data to develop knowledge within organisations (Tadejko, 2020). Access to these cloud-based solutions unlocks powerful computational capabilities, which enable organisations to develop and apply complex AI models across various business processes.

4.3.4 AI-based training systems for personalised employee development

AI-based training systems support personalised employee development by analysing initial skills, cognitive levels, job requirements and organisational needs. These systems learn from behavioural data to refine employee profiles and create tailored learning strategies (Maity, 2019). With the use of NLP, virtual trainers can extract content from company training materials and adapt it to individual learning preferences, enabling customised training. Furthermore, AI-based trainers can act as personal mentors, identifying skill gaps and providing targeted support in interactive learning environments. This reduces human interventions and bias, while facilitating efficient, learner-centred training (Maity, 2019). Additionally, these systems support training needs analysis and apply algorithms to provide timely repetition of knowledge at moments when users are most likely to forget (Owoc, Sawicka and Weichbroth, 2021).

From another perspective, AI-based virtual trainers have the potential to analyse and measure employee performance against pre-defined parameters. This supports the *knowledge measurement* process to assess the effectiveness of KM initiatives. The application of AI-based virtual trainers could revolutionise individual learning through mass personalisation and individual knowledge development.

4.3.5 Enhancement of creativity and innovation

Creativity is essential for innovation, problem solving and progress. While Probst, Raub and Romhardt (2012) have traditionally viewed creativity as spontaneous and uncontrollable, AI technologies demonstrate that creativity can be enhanced with technological support. Generative AI models have made significant progress in the development of knowledge. ChatGPT for example, utilise large, diverse datasets to generate coherent, context-aware responses (Korzynski et al., 2023). This opens up innovative avenues for idea generation. ChatGPT can be used at various stages of the innovation process, including idea generation, brainstorming and the design process. The text-to-code generation feature allows the creation of code, images, user journeys and prototypes using natural language, which supports creativity and the development of ideas (Bilgram and Laarmann, 2023). This capability also enables less technically skilled users to create early digital prototypes

that can be converted into program code. Another example is a prototype presented by Botega and da Silva (2020), which uses AI to support KM in the selection of methods that foster creativity and innovation in the design process. To overcome creative blocks and promote the development of creative solutions, this AI-based system stores, retrieves and applies knowledge about creativity and innovation methods.

4.3.6 The role of chatbots in modern KM

In the rapidly evolving AI landscape, chatbots in particular play an important role in modern KM. These technologies not only streamline information retrieval, but also foster a more open and collaborative knowledge culture within organisations. Chatbots provide an informal and accessible way for employees to request information without the fear of being perceived as incompetent by other employees (Jarrahi et al., 2023). ChatGPT, for example, offers a simple conversational mode that provides personalised answers to open-ended questions and adapts to the user's language. This user-centred and user-friendly adaptation facilitates and secures organisational knowledge retrieval (Korzynski et al., 2023). Chatbots broaden the knowledge spectrum of users by providing explicit knowledge through general query capabilities (Ali et al., 2023). ChatGPT can be used for a variety of NLP tasks, such as the creation and translation of scripts and documents (Ali et al., 2023). Additionally, ChatGPT can analyse and summarise both unstructured and structured data, making it easier and quicker for decision-makers to find the information they need (Korzynski et al., 2023).

4.4 Four-Stage Model to Support the Creation of Best Practices and Lessons Learned

Probst, Raub and Romhardt (2012) emphasise the importance to document team experiences during projects. This can be valuable for teams that face similar challenges. However, experiences are often not systematically documented at the end of projects.

Given that no specific AI technologies were identified in the literature for this area, this study proposes a four-stage concept. The concept is based on existing knowledge to help organisations systematically document lessons learned and best practices. Figure 3 shows the four phases of this concept.

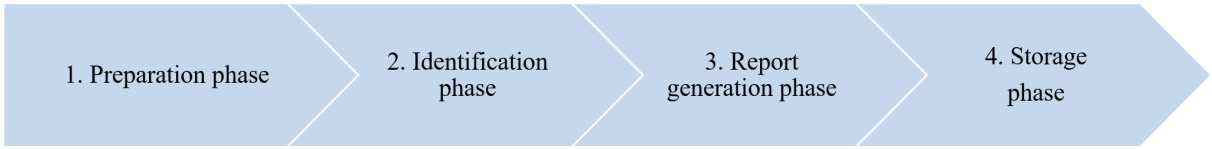


Figure 3: Four-stage concept for best practices and lessons learned (Source: Figure created by author)

In the *preparation phase*, AI technologies such as ML, intelligent agents and NLP are recommended. These technologies enable the effective capture and analysis of extensive data from various sources. This includes project documentation, emails, meeting minutes and performance reports. Studies by Hilger and Wahl (2022), Maity (2019) and Masood and Hashimi (2019) demonstrate that these AI technologies are particularly well-suited to the extraction and consolidation of relevant information from large datasets. This capability is crucial in the *preparation phase* as it provides a solid foundation for the subsequent phases.

In the *identification phase*, AI technologies such as ML and NLP are recommended to analyse the data collected in the first phase and identify crucial patterns and insights. ML excels in the recognition of recurring problems and trends, while NLP specialises in the extraction of key themes from text-based data. These technologies provide deeper insights beyond what human analysts could capture. This makes them essential to identify effective strategies, patterns, relationships and potential challenges. Research studies by Hilger and Wahl (2022) and Masood and Hashmi (2019) emphasise the effectiveness of ML and NLP to uncover complex data patterns.

In the *report generation phase*, natural language generation (NLG) is recommended to translate insights from analysis and pattern recognition into automated, easily understandable reports. NLG can convert complex insights, data and patterns from projects into text form. These reports can be customised to the specific needs of the organisation. NLG translates machine actions into language that is comprehensible for humans, which is crucial to effectively communicate analysis results (Paschen, Kietzmann and Kietzmann, 2019).

In the *storage phase*, the implementation of a knowledge database is recommended. This database serves to store all lessons learned and systematically categorise them. Equipped with an intelligent search function, the knowledge base provides users with quick access to relevant information. Hilger and Wahl (2022) emphasise the benefits of intelligent search functions.

Alternatively, the proposed concept to document lessons learned and best practices can be implemented effectively in a chatbot. The conversational mode can increase user-friendliness and enhance employees' willingness to interact with the system (Jarrahi et al., 2023). Interestingly, ChatGPT already provides the capability to develop custom chatbots. This makes it a valuable resource to implement this concept. However, the use of ChatGPT must comply with the data protection requirements and regulations of the organisation.

4.5 Changes in KM Through AI: The Sixth Generation of KM

The examples in Section 4.3 illustrate various AI technologies and their applications within KM processes, emphasising the diversity and rapid development of AI in this field. These advancements illustrate the dual potential of AI: it can support individual KM processes and create synergies as part of a comprehensive system to improve overall efficiency. AI technologies offer tailored solutions that streamline traditional KM approaches and enhance process effectiveness.

A critical review of the literature reveals that most studies concentrate on specific KM aspects rather than the entire process. However, this review confirms the versatile applicability of AI technologies across seven KM building blocks, thereby broadening the understanding of its potential in KM. The comprehensive integration of AI could unlock new perspectives and possibilities. The current analysis highlights the deployment of NLP and chatbots across five building blocks, which serves to illustrate their flexibility, central role and transformative impact in modern KM. This trend highlights the need for organisations to invest in these technologies and adapt their strategies accordingly.

As introduced in the initial chapter, Bencsik's (2021) study provides a foundation to explore the changes in KM due to AI. Bencsik's analysis of the historical evolution of KM over five generations provides a contextual framework to forecast the next generation of KM. To address the research gap presented in the introduction (Chapter 1), the sixth generation of KM required an examination of both *evolutionary* and *revolutionary* approaches. At first glance, the integration of AI into existing KM systems appears to be *evolutionary*, as it often supports established KM processes. However, a more detailed examination reveals that AI has the potential to achieve a *revolutionary* level of advancement. The capabilities presented in Section 4.3 – including generative models and technologies such as ChatGPT – demonstrate that AI can drive KM beyond traditional support functions to generate and process knowledge in unprecedented ways. AI technologies that go beyond pure support functions – such as the automatic generation of knowledge from large databases and the development of self-learning, adaptive systems – allow for a fundamental redesign of KM processes. Therefore, AI opens up a wide range of innovative applications. These range from the enhancement of the interaction and learning experience to the promotion of creativity and the support of complex decision-making processes. Such technologies extend far beyond mere integration into existing systems and revolutionise the way knowledge is created, distributed, used and preserved.

This study suggests that the sixth generation of KM, as envisioned by Bencsik (2021), may indeed follow a revolutionary path. The integration of AI into KM represents a strategic reorientation towards the development of integrated and intelligent KM.

4.6 Impact of Chatbots on KM Building Blocks

This study highlights the use of chatbots in five KM processes, as detailed in Table 2. Chatbots contribute to improvements in efficiency and in the promotion of innovation in organisations. The focus of this section is to identify which KM building block is most affected by chatbots. This can be achieved through the analysis of the concept matrix presented in Table 2. The concept matrix analysis reveals that chatbots significantly impact the *knowledge use* process, with four visual “X” marks. The area of *knowledge distribution* also shows notable integration, with three marks. There are two marks for *knowledge development* and one mark each for *knowledge acquisition* and *measurement*. This indicates a smaller but present influence.

In summary, the concept matrix confirms the transformative role of chatbots, particularly in the process of *knowledge use*. Their broad applicability underscores their potential to enhance various KM processes, as discussed in Subsections 4.3.5 and 4.3.6. This study highlights the importance of the integration of chatbots into KM processes in order to realise their full potential for efficiency improvement and innovation.

5. Conclusion

The integration of AI in KM enables organisations to access the required knowledge for more efficient decision-making. AI has the potential to significantly increase efficiency while augmenting human tasks. The sixth generation of KM appears to be a *revolutionary* approach to the transformative impact of AI

technologies. It is characterised by the deep applicability and the extension of AI in KM beyond the pure support functions. This integration not only redefines traditional KM processes, but also significantly enhances the use of knowledge in ways never before imagined. This systematic review aims to be a valuable resource for scholars and practitioners. It provides insights into the opportunities associated with AI technologies and their symbiotic relationship with KM processes. The result showed that the most frequent AI technologies include NLP and chatbots. These technologies proved crucial in five identified building blocks in the literature. The results also indicated a significant relationship between *knowledge use* and chatbots. This improvement leads to a more dynamic and efficient use of knowledge.

This research has successfully achieved its intended objectives to explore the integration of AI technologies in KM processes. However, several limitations must be acknowledged and addressed. The focus on the eight specific databases may have inadvertently excluded potentially valuable publications. Additionally, the selection of search terms may have influenced the study's scope. Future studies could refine these parameters for broader inclusivity and precision. Furthermore, the focus on literature from 2019 to 2023 may have overlooked earlier foundational research in AI and KM.

This study identifies avenues for future research to improve our understanding and application of AI in KM. There is a notable gap in the use of AI to define and formulate *knowledge goals*. This suggests the need for further research into how AI could support or augment the human judgement required to set organisational goals. Additionally, future research should focus to evaluate the efficiency of specific AI technologies in real-world KM contexts. The effective exploitation of the transformative potential of AI will be crucial. Organisations seek to revolutionise knowledge processes and take advantage of ongoing technological advances.

Acknowledgements

The author would like to thank Dan Remenyi for the opportunity to contribute to the conference and the ICAIR Conference Committee for their efforts in organising this event.

References

Abubakar, A. M., Behravesh, E., Rezapouraghdam, H. and Yildiz, S. B. (2019) “Applying artificial intelligence technique to predict knowledge hiding behavior”, International Journal of Information Management, Vol 49, pp 45–57.

Ali, O., Murray, P., Momin, M. and Al-Anzi, F. S. (2023) “The knowledge and innovation challenges of chatgpt: a scoping review”, Technology in Society, Vol 75, pp 102402.

Bencsik, A. (2021) “The sixth generation of knowledge management–the headway of artificial intelligence”, Journal of International Studies, Vol 14, No. 2, pp 84–101.

Bilgram, V. and Laarmann, F. (2023) “Accelerating innovation with generative ai: Ai augmented digital prototyping and innovation methods”, IEEE Engineering Management Review, Vol 51, No. 2, pp 18-25.

Botega, L. and da Silva, J. C. (2020) “An artificial intelligence approach to support knowledge management on the selection of creativity and innovation techniques”, Journal of Knowledge Management, Vol 24, No. 5, pp 1107–1130.

Brocke, J. v., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R. and Cleven, A. (2009) “Reconstructing the giant: On the importance of rigour in documenting the literature search process”.

Davenport, T. H. and Prusak, L. (1998) Working knowledge: How organizations manage what they know. Harvard Business Press.

Hilger, J. and Wahl, Z. (2022) Text analytics tools. In Making Knowledge Management Clickable: Knowledge Management Systems Strategy, Design, and Implementation, Springer, pp 193–197.

Jarrahi, M. H., Askay, D., Eshraghi, A. and Smith, P. (2023) Artificial intelligence and knowledge management: A partnership between human and ai. Business Horizons, Vol 66, No. 1, pp 87–99.

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K. and Ziemba, E. (2023) “Generative artificial intelligence as a new context for management theories: analysis of chatgpt”, Central European Management Journal, Vol 31, No. 1, pp 3-13.

Maity, S. (2019) “Identifying opportunities for artificial intelligence in the evolution of training and development practices”, Journal of Management Development, Vol 38, No. 8, pp 651 663.

Majumder, S. and Dey, N. (2022) AI-empowered Knowledge Management, Springer, Singapore.

Masood, A., Hashmi, A., Masood, A. and Hashmi, A. (2019) Knowledge management & intelligent search. Cognitive Computing Recipes: Artificial Intelligence Solutions Using Microsoft Cognitive Services and TensorFlow, Apress.

Mohammad, A., Zahrawi, M., Al-Emran, M. and Shaalan, K. (2022) “A systematic review on the relationship between artificial intelligence techniques and knowledge management processes”, in: Al-Emran, M., Shaalan, K. (eds) Recent Innovations in Artificial Intelligence and Smart Applications, Vol 1061. Cham: Springer, pp 67–83.

Nonaka, I. & Takeuchi, H. (1995). The Knowledge-creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, New York.

Owoc, M. L., Sawicka, A. and Weichbroth, P. (2019). 'Artificial intelligence technologies in education: benefits, challenges and strategies of implementation'. Artificial Intelligence for Knowledge Management, AI4KM 2019. Macao, China, 11 August. Cham: Springer, pp 37–58.

Paschen, J., Kietzmann, J. and Kietzmann, T. C. (2019) “Artificial intelligence (ai) and its implications for market knowledge in b2b marketing”, Journal of business & industrial marketing, Vol 34, No. 7, pp 1410–1419.

Polanyi, M. (1966) The Tacit Dimension. Doubleday.

Probst, G., Raub, S. and Romhardt, K. (2012) Wissen managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen, Vol 7. Gabler Verlag, Wiesbaden.

Tadejko, P. (2020) 'Cloud cognitive services based on machine learning methods in architecture of modern knowledge management solutions', in Poniszewska-Marańda, A., Kryvinska, N., Jarząbek, S., Madeyski, L. (eds.) Data-Centric Business and Applications: Towards Software Development, Vol 4. Cham: Springer, pp 169–190.

Tamla, P., Hartmann, B., Nguyen, N., Kramer, C., Freund, F. and Hemmje, M. (2023) 'CIE: A Cloud-Based Information Extraction System for Named Entity Recognition in AWS, Azure, and Medical Domain', Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2022. Valletta, Malta, 24-26 October. Cham: Springer, pp 127–148.

Turing, A. M. (1950) Computing machinery and intelligence, Oxford University Press.

Vajpayee, A. and Ramachandran, K. (2019) “Reconnoitring artificial intelligence in knowledge management”, International Journal of Innovative Technology and Exploring Engineering, Vol 8, No. 7C, pp 114–117.

Webster, J. and Watson, R. T. (2002) “Analyzing the past to prepare for the future: Writing a literature review”, MIS Quarterly, Vol 26, No. 2, pp xiii–xxiii.

Artificial Intelligence and its Role in Legislative Practices

Marius Vacarelu
National School of Political and Administrative Studies, Bucharest, Romania

marius333vacarelu@gmail.com

Abstract: One of the most significant challenges faced by the average person is understanding the general context of the legislation that applies to them. The legal discipline has certain characteristics that make it almost esoteric for those who are not part of it, however, it is necessary to know the legal norms as comprehensively as possible. People today are no longer just "from a village/town/province/country", but have become something almost universal within the frameworks of globalization and the vast library the Internet offers at nearly zero cost. Artificial intelligence (AI) is a technology that is beginning to fundamentally change society, and within this "sea of transformations", the law – and legal and political practices in general – cannot avoid contact and change. There is no area of law not being altered, but in my opinion, the most significant place where transformations will be recorded is in legislating and drafting normative acts. Legislative operations are always complex and rarely bring satisfaction to those subject to regulation, given the relationship between the rights and obligations set out by normative acts. At the same time, it is challenging to legislate in an increasingly complex society, where enduring and situational interests intersect, where there are poor-quality mechanisms in legal documentation, and where the concept of legality is not always correctly perceived in the political environment. AI should offer a greater understanding of legal concepts to both ordinary citizens and legislators, precisely through its extensive library and its demonstrated synthesis and analysis capabilities. Therefore, the use of AI in the legislative process will become increasingly necessary as its capabilities grow, primarily to produce faster syntheses of legal documentation, essential for correctly understanding the context that justifies the need for new legislation. I believe that in the coming years, no country will escape this change, and AI will help eliminate some of the poor-quality practices that do not offer countries and citizens better prospects for wealth and professional development.

Keywords: Legislation, Artificial intelligence, Transformations, Documentary, Codification

1. Introduction in General Theory of State and Regulation

In a state that respects the rule of law, no one is above the law, not even the Government. It is, therefore, crucial that checks and balances are embedded in the constitution to prevent the executive from exerting undue control. The essence of legality is that action by the state and its agents is both in accordance with and authorised by the law (JUSTICE, 2023, p. 10). For this reason, neither the use of Artificial Intelligence in the entire legislative process nor the application of normative acts can violate the principles of law, if it is desired that the state and society maintain the balance that underpins the economic and personal development of citizens, public institutions, and businesses. Even though this new electronic tool offers broad perspectives for action and the enforcement of one's own interests – both for governments and other actors with political, legal, or financial power – it will still be necessary to uphold the sacrosanct principles of the rule of law, in order to prevent major regressions that could ultimately lead to social upheavals or even reckless geopolitical actions.

The essence of the concept of "rule of law" is straightforward: it involves the adoption of legislative acts that are accepted by absolutely all individuals within a country, without any special or generalised immunity. Explained in this way, all players in the social, economic and political arena will adhere to the norms issued by various centers, whether national or local, based on their material and territorial jurisdiction (Bingham, 2010, pp. 27 – 28). Compliance with legal norms should be conducted – at least theoretically – without any hesitation or questioning their text (Barnett, 2002, p. 74). It is, therefore, presumed that legal norms are always very clear, very well balanced, and, most importantly, can be amended at any moment deemed appropriate, following simple procedures known to each citizen (Barraud, 2016, p. 6). Obviously, public administration also has the same obligation in its process of organizing the enforcement of laws and in terms of their concrete implementation. However, public administration institutions will face the practical effects of AI sooner than parliaments, and therefore they will be the first to need to alert lawmakers regarding the changes that will need to be constantly made to the normative acts that will regulate the use of this technology.

Nonetheless, reality and basic human caution necessitate a different approach, as there are two issues regarding the quality of regulation that must always be considered and everywhere: what is the purpose of the regulations (in plain language, who benefits the most from them), and what are the procedures for adopting and amending these regulations.

For sure, any person will understand from the outset the necessity of comprehending the real purposes of a normative act, as it can be expressed directly and unequivocally. The purpose, as the general theory of legal acts states, is an essential element of human will, without which these acts would not be valid, and thus it is easier to detect (Millard, 2022, pp. 24 – 26). Since each of us pursues a specific purpose in our actions, it is easy to understand the intrinsic link between legislation and the objective that political power aims to achieve through the adoption of these legal norms.

However, procedural aspects are of greater importance. Often forgotten or overlooked, they are what give substance to the adoption of normative acts, and most debates arise precisely in relation to the types of procedures used either for adoption or for modifying certain essential aspects of legislation. Most of the time, these are "under the media's radar", which does not always understand them very well and requires detailed explanations to provide the public with a genuine understanding of what is happening in the plenary halls of parliaments and governments. Nevertheless, due to their complexity and subtlety, these procedures offer the occasion for lively intellectual debates, wherein the entire capacity for argumentation is weighed to persuade the majorities in political-administrative bodies to adopt various legislative measures (Keesey, 2018, pp. 41 – 42). This text attempts to briefly analyze – due to space constraints imposed by the editor – some issues related to the relationship between legislation adopted by parliaments and Artificial Intelligence. If we were to extend the analysis to the possibilities of action for public administration and the relationship it will develop with AI, we would quickly arrive at a minimum 100-page essay, which compels us to show respect and self-restraint towards the reader: the depth of the author's thought is not lacking, but it cannot always be expressed in a few sentences, rather in hundreds of very long paragraphs.

Once again, it is necessary to emphasise that legislative procedures are viewed more from the perspective of the steps to follow and the various voting operations, including delaying the vote (when the minority lacks voting power but has sufficiently plausible arguments to prevent a swift vote). Generally, this includes matters related to the quorum for sessions or voting. Without contradicting this aspect – for which there is ample practice, confirmed by my personal experience – we must highlight that the pre-voting stage of the legislative procedure is more significant.

What comes before voting procedures within the legislative process? The set of operations and preliminary procedures leading up to the votes can be summarised under two main headings: the necessary documentation for adopting or amending (including repealing) normative acts, and the drafting of the specific act itself.

As we shall see, in both scenarios the capabilities that Artificial Intelligence (AI) offers are – and will certainly be – used to create normative products that, at least theoretically, should be very good. In fact, if we closely analyse a significant portion of what today's AI providers offer their clients, we will notice that they largely fit into both options, with some results that have already produced certain legal consequences (Sanders and Schneier, 2023).

2. Legislate Process' Principles

Legal theory in the realm of legislation highlights that the process of drafting legal norms must be supported by data, information, and the conclusions of scientific research. This rule, found in almost all legislations is based on the idea that the legislator must draft normative acts based on preliminary economic, sociological, and psychological investigations and studies, scientific documentation, and specialized research in the field where the regulation is taking place, in order to thoroughly understand reality and identify the goals of legislative policy. The proposed solutions must be scientifically grounded and argued, and when multiple hypotheses can be considered for resolving issues, several possible alternatives will be presented. The scientific basis for a law must mandatorily include a description of the actual situations to be transformed into legal situations; an analysis of the value judgments regarding the determination of the actual situations that need to be transformed, changed, or supplemented, which are in contact with the value judgments; the social cost of the legislative reform and its appropriateness. Scientific research must also underpin some short, medium, and long-term legislative forecasts and eliminate opportunistic or inadequately researched legislative actions. Comparative research of the legislation from other states can also be utilised (Marinese, 2008).

As can easily be observed, it is straightforward to identify the main characteristic that a normative act must possess, namely the scientific approach to an issue that needs regulation. Hence, we are dealing with a matter of almost mathematical nature, where all hypotheses of a situation need to be precisely identified, along with all elements that have contributed to its emergence, as well as all the consequences it entails both in the short

and long term. In fact, an essential aspect of a normative act is its ability to remain effective over the long term, sometimes for decades or even longer, which means that the legislator, at the time of drafting the text, had a profound capacity for synthesis and long-term thinking (Chevallier, 2015). The principles of law which remain the most enduring over time, especially those in the realm of private law, are normative acts that are written/adopted in violation of these principles. They cause problems in practice from the moment they come into force, requiring subsequent modifications.

3. Artificial Intelligence and its Legal Potential

In this broad description, what would be the role of Artificial Intelligence? Firstly, it is a relatively new technology, created with the aim of providing intellectual support to humans – although, when installed on mechanical devices, it can very well take on tasks of a physical nature, as a recent news report about replacing 100,000 jobs with 750,000 robots demonstrates (The Business Standard 2024).

If we carefully analyse the concept underpinning the entire idea of "artificial" intelligence, we find humanity's aspiration for perfection. In fact, the entire history of literature and human myths centres on the idea of humans becoming perfect, or being able to achieve perfection, with the aid of a magical potion, a wondrous/enchanted instrument (sword, ring, etc.), an older person who provided good advice, or very often, the invaluable support of a horse.

The value of the advice given by a wise elder or that of a horse – both beings endowed with a soul – is more valuable to the hero than various enchanted objects. In the same paradigm, artificial "intelligence," as it is, seeks to rise to the level of a "soul", capable of understanding, surpassing the state of an object, which is handicapped by the absence of this soulful element (Campbell, 2008). Although this might seem like a digression from the subject of this study, the truth is that the full scope of Artificial Intelligence cannot be understood without grasping the reasons for its emergence, even in legislative process. Humanity wanted to fly, so it invented the aeroplane; it wanted to jump from the air, so it invented the parachute; it wanted to dive hundreds of meters underwater, so it invented diving gear. Similarly, Artificial Intelligence emerged as an assisting tool akin to those in stories, in the guise of "wisdom", enabling a person's cognitive and work capabilities to reach a level beyond physical limitations. As we can easily observe, the most significant contribution Artificial Intelligence makes is of an individual nature, allowing its operator to obtain benefits both for themselves and for the community they lead or serve (as happens in various fairy tales where slaying a dragon liberates an entire country).

Legal acts always have an initiator – at least one person, who, regardless of the title given in legislation, will also draft the bill. And draft laws must also – at least theoretically – aim for the highest quality in terms of regulation, long-term projection of legal situations, based on the analysis of various data sets available to the initiator of the normative act (Karpen, 2019). In practice, researching the issue that the initiator seeks to regulate primarily falls on them, and in this regard, they must collect data from various sources, both available on the Internet and in printed formats or recorded on media that have not been digitised. The time frame that legislators have for various bills must also be taken into account, as sometimes it is compressed into a period of just a few days, while in other situations there is no real pressure from deadlines. At the same time, ignoring the aspect of party interests makes the analysis of legislation superficial – as a result of these interests, some bills advance through parliamentary procedures very quickly, while others can be delayed for months or even years, which affects not only the entire national legislative framework but also the idea of adequate documentation, as each day brings some changes to human society.

Collecting the necessary data for drafting a regulatory act is a challenging, laborious operation that requires a significant amount of time and always raises a problem: which data are more relevant than others? In a period when the Internet itself offers an "ocean of information", it results that the collection operation is difficult, and the human mind struggles to process the obtained data. This is precisely where Artificial Intelligence can intervene – and it will undoubtedly do so increasingly in the coming years.

AI-powered tools and algorithms now have the capability to gather, process, and interpret massive volumes of data at an unprecedented speed and accuracy. Through advanced algorithms, AI can collect data from diverse sources, including IoT devices, social media platforms, and online interactions. This enables public servants and dignitaries to access real-time, high-quality data, leading to more informed decision-making. Moreover, AI has facilitated the integration of structured and unstructured data, allowing public institutions to gain a holistic view of their administrative behavior and citizens (Stagnate Research 2023).

As mentioned above, an important element in the necessary documentation for adopting a legal act is contact with the legislation of other countries. Typically, this was only accessible in very small proportions before the development of the Internet, and what was available usually belonged to just a few major countries. The emergence of the Internet has provided a large dose of available legal acts in multiple languages, but there is a problem: the general understanding of the texts, which – considering the legal traditions of each country, as well as their linguistic peculiarities – are difficult to translate accurately, at least by non-professional high-level translators. The advent of electronic translations via browsers has alleviated part of the problem (Hutchins 2003), but not as much as lawyers drafting the legal texts would like, and even less so for politicians who genuinely wish to contribute to the development of their own countries.

However, with recent developments, we can observe that transcription tools powered by AI, equipped with sophisticated speech recognition algorithms make the process of transcription of qualitative data tasks significantly easier for researchers since such tools can transcribe data by automatically converting audio or video recordings into written text (Christou 2023).

In practice, Artificial Intelligence is capable of producing competent documentation based on the electronic sources available on the Internet. We could make a reasonable observation here regarding the quality of regulation: in countries where archives are becoming increasingly digitised, documentation can be more extensive, with a quality level that should be maximal.

Unfortunately, as judicial practice has shown, there is also the risk that some AI providers might create situations that fit a particular argumentation. Thus, only in 2023 we have the 55% of US lawyers that they use AI for case documentation (Juro 2023); likewise, British judges are permitted to use this technology in their reasoning of decisions (The UK Judicial Office 2023), and there are also instances where ChatGPT invents certain arguments (cases with the value of judicial precedent) to support a lawyer's argumentation in a trial (Case 22-cv-1461 2023).

The issue of documenting jurisprudential aspects is not specific to the common law system; it is significant in any legal system, and AI can greatly assist those trying to better understand the process of drafting – but particularly of amending – a legislative act. In practice, legislators often assume that the impact of the laws they enact on everyday life is purely positive, while negative aspects are considered accidental or possibly the result of misapplication of the law.

Therefore, it is generally difficult to convince a legislator – government or parliamentary party (or a member of these types of institutions) – that their legislative act is fundamentally flawed from the outset (Enright 2019), that it primarily causes disruption in people's lives, and that the negative and unwanted effects will be visible almost immediately. The psychology of a politician in a parliament should therefore not be ignored, as they act both convinced of the legislative power they possess and constrained by one of the most important interests of any party: that of not being seen as the author of any errors (political or otherwise) that could be punished by the electorate in the next elections (Weinberg 2011).

The role of Artificial Intelligence in the documentation work necessary for drafting normative acts would involve a broad analysis – but primarily the collection – of as many judicially deduced situations as possible. This is crucial because a person ends up in court when other administrative or mediation procedures have failed, indicating high levels of dissatisfaction and a lack of understanding from those enforcing the legislation perceived as inadequate by citizens. Practically, AI could better identify – with its synthesis capacity – the most common behaviours leading to court actions, the types of actions exhibited by various administrative actors (Pszczyński 2021), and whether there were alternative, non-judicial remedies available before resorting to the courts. Furthermore, AI should extract from court decision rationales what is considered by the judiciary as violations – particularly frequent ones – of laws and general legal principles.

Thus, what AI should do in the field of documentation is to inventory the errors that the existing laws have produced, highlighting the most common issues encountered in practice. Obviously, both normal human matters that continue to cause significant harm, such as the egos of those who initiated a legislative project or the financial interests of the initiators or their sponsors, must be taken into account. Additionally, AI will not be able to consider two other practical aspects: a) the ideology of different political parties, which can sometimes lead to an ideological deadlock – how to adopt a measure that conflicts with the general trends of the party? and b) classic political interests, whereby the price of certain political coalitions involves amending some normative acts based on the pragmatic interests of the parties involved, along with the interests of their sponsors. Artificial Intelligence is designed to be as neutral as possible in relation to facts and ideologies;

however, common practices reveal that human interests can contradict legal principles, and ultimately, societal dissatisfaction can be widespread, even though the main political leaders (and their main sponsors) may be satisfied.

The AI's capacity for documentation would thus be useful in terms of identifying judicial practice aspects, but it would not be sufficient. A legislative act produces consequences from many angles, and not all necessarily end up in court. Therefore, a legislative act can be flawed in terms of its ideas, without violating legal drafting principles – for instance, a fiscal policy that increases taxes in a specific sector, which then begins to contract. In this scenario, the data that needs to be corroborated comes from a different domain and not necessarily a national one since certain investors might relocate to other countries, and the data could come from analysing sources published abroad – such as financial reports of investors driven away by the tax increases, which appear a few months after the respective company has withdrawn from the country/market.

An important role that AI can play in the documentation process prior to the development of a legislative act is monitoring social media. As is well known today, especially from examples in the political competition realm, from Cambridge Analytica to its successors (Westby 2019), it is very helpful to know society's opinions, however pleasant or not they may be. For a legislative act, it would therefore be very useful to see in advance what the general reactions of society are, how influential individuals position themselves in the analysis of various legal provisions, and thus you can anticipate some of the possible media debates or even potential legal actions. Let us not forget that history has shown how sometimes a single piece of legislation was able to trigger social upheavals, following which public institutions were attacked by furious populations, or regime changes took place. Evidently, we are also very aware of the flip side of the coin, where AI can – and unfortunately is frequently used to – monitor social media to preempt any type of protest (Brumfiel 2023). However, conducting a scientific analysis of a situation and a legislative project, if done properly and in good time, could have prevented a significant part of social disturbances or unwanted economic and fiscal effects.

In essence, we return to what I mentioned earlier: this new intelligence seeks to be – as least in its projection – as close as possible to a scientific truth, striving to remain neutral and, based on the available data, to be as useful as possible to those aiming to achieve a significant goal, such as perfecting the regulatory act in the field of legislative drafting.

Scientific opinions today are no longer solely the preserve of universities and various research institutes. The universality of social networks – today, the most popular of which boasts over 2 billion daily users (McLachlan 2024) – means that the voices of many professionals are heard and can influence political as well as administrative behaviour. Without delving into the mechanisms of social influence, as our text serves a different purpose, it is important to highlight that professionals in any field of the economy can quickly see the impact of a new legislative act (whether a proposal for a new act or a modification of an existing one). It is mandatory to consider the opinions of those who are well-versed in a particular sector, if only to avoid the phenomenon known as brain drain, which is unacceptable in the 21st century. Thus, AI can assist in gathering various opinions of professionals from both past and present, to provide lawmakers with a clearer picture. We must remember that legislative processes can often take months rather than weeks, or even years, especially if the issue being addressed by the parliaments is morally, politically, economically, socially, or psychologically complex.

Good documentation in any legislative effort requires knowledge of all the existing regulations that govern a specific area. This is logical because new regulations do not emerge in a "vacuum"; particularly in the 21st century, they are often amendments to other texts. This happens either because new situations have arisen that were not covered by previous regulations or because certain political, economic, or lobbying interests demand a different regulation. In any case, the actual legislation must find a way to harmonise the new regulation's content with what already exists, to ensure no gaps in regulation, at least minimally, and to avoid accidentally repealing useful provisions contained in various regulations. Therefore, a systematic analysis is needed to identify all the regulations governing an area, any overlaps, and especially their legal force, as an act passed by a parliament is stronger than one adopted by a local council or a ministry. This task is very difficult in practice because many regulations contain brief provisions regulating specific aspects of a domain generally governed by another regulation. This situation is called a "strict interpretation and application exception," and where there is an abundance of regulations, it can arise easily but without coordination.

One purpose of regulation, if done according to the norms of legislative science, is to systematise a domain of activity's regulation into a single act as much as possible, to facilitate a nearly complete understanding of it (Padjen 2019).

Only after thorough documentation has been conducted can the legislate process begin, specifically the writing of the legislative text. In this scenario, Artificial Intelligence can be useful again, as it can quickly generate various working drafts, even in language that is not necessarily legal. In fact, legal science emphasises that the regulatory framework can be written in terms more common to other fields, provided it clearly specifies what needs to be regulated and how this should be done, in broad terms. In this regard, it should not be forgotten that not every member of a parliament is a graduate of legal studies, and therefore their assistants must first convince them of the correctness of the ideas contained in a draft law that they need to initiate. The absence of any educational requirements to run for office may mean that the assistants of MPs can be chosen not always based on high competence, and this lack of legal knowledge can manifest in any form, with or without written justifications by a human or by an AI tool guided by poorly trained professionals. Thus, Artificial Intelligence will not be a panacea for drafting normative acts; it could even be a threat, as in the legislative process it can be used by political opponents for analyses that highlight each other's shortcomings in various draft laws.

The advancements that Artificial Intelligence has reached today have even enabled the drafting of quite coherent legal documents. By the end of 2023, the first legal document entirely written by this new technology was already adopted, taking only 15 seconds, in Recife, Brazil (Buttice 2023).

In other countries, however, AI has not yet reached that level of acceptance by lawmakers (Hill 2023), partly because the complexity of legal systems varies, and the expectations for regulation are always high. Nevertheless, it can be noted that at the level of local public administration and even across various institutions, it will be easier to envisage and even implement lower-level regulations compared to those adopted by parliaments or governments. This is because these regulations are less complex and – importantly – there is not always the same calibre of legal expertise within these institutions.

In countries where Artificial Intelligence can be used to its fullest potential, however, there is no equality in the financial power of political actors. In the legislative process, we will find this technology at various stages. First, it will be important in the documentation phase, including by providing translations from languages other than the native language of the initiators of new laws; it will be useful in drafting different text variants, although the version that will be made public will depend on human political will. Within these text formulations, this technology can perform simulations regarding the effects of different emerging variants, highlighting vulnerabilities and risks. Subsequently, in the public phase of legislation, AI can be used by various political opponents or by citizens or private agents for critical analyses, which can lead – whether with or without the assistance of this technology – to the formulation of proposals to amend the initially proposed text. Meanwhile, AI can be used to monitor societal reactions to the draft law, and based on the various aspects that will be revealed, the project may be modified or even withdrawn. Similarly, AI can also be used after the adoption of the regulatory act itself, and based on the reactions observed in administrative practice and the economy, further modifications to the text of the law can be made, or even a new regulatory act can be initiated to better regulate the situation, with the initial act being repealed.

Through the lens of importance, we will make a brief note about how local public administration can use AI in the process of enacting its own legal norms – with legal power inferior to the acts issued by parliaments, but with substantial relevance and applicability in practice. Thus, it may be that in a local public administration institution within a smaller and less affluent community, the conception and drafting of legal acts with the assistance of Artificial Intelligence could be a better and more cost-effective solution. This is because a poorly trained jurist could make significant errors in drafting such acts if they do not have a specialist nearby to assist them with legislative matters. Practically speaking, as mentioned earlier, in such cases, AI would play a role similar to that of the horse that advised and partially resolved the challenges faced by the hero in fairy tales, before the hero eventually became a king or a prince consort and future king.

We do not claim that the direct effect of increasing artificial "intellectual" capabilities will be the full drafting of normative acts. In fact, knowing – as a jurist – the thinking process of the profession, I am aware that this is not desired, and efforts will be made to prevent the promotion of AI in the sphere of drafting normative acts. At the same time, it is equally evident that, at least in the documentation necessary for adopting normative acts, AI already has a firmly established role (within less than three years), and this will increase, eventually taking over part of the text drafting process – from short regulations to big codes, especially in commercial area. However, what AI has already triggered in the realm of regulation is the increase in debates within institutions involved in legislation, as AI provides multiple text variants, but aligning them with the pragmatic interests of individuals can only be done by those who are directly interested and/or affected.

4. Conclusions

The very idea of regulating society has always been a prerogative of power, which has sought to organise communities, but at the same time to subordinate them. The struggles for political power have always involved this prize, as legal norms regulated the relationships between institutions and the permissions for different political or political-economic factions. However, normative acts always had to adopt a scientific form, outlined by the Science of Law, so that even the most oppressive dictatorship was obliged to give a legal façade to any idea, no matter how abusive or absurd it might have been.

The technological development of recent decades has transformed society, which has become more multi-vocal, and the state has had to adapt to this expansion of social involvement. The emergence and development of AI have enabled not only political operations but also the enhancement of all informational possibilities for both public institutions and citizens regarding good legal and administrative practices, which are intended to be replicated through normative acts. The AI's ability to document as thoroughly as possible allows for an increase in the quality of regulatory acts, both at the central (governmental and parliamentary) and local levels. Evidently, if AI has provided good results in documentation, it has logically led to the desire to extend its capabilities to the drafting and writing of normative acts, an aspect which has not yet been achieved to the same extent.

It will be interesting to see in the coming years what the outcome of the progress in Artificial Intelligence technology will be in the field of legislation. As is true in any country, legal acts are what organise human life. A failure of AI in this matter would only serve to increase citizens' distrust in political systems, which would be blamed for the failures of the technology. However, we believe that a *modus vivendi* can be reached between the minds of lawyers, the minds and interests of politicians, and the capabilities of AI. Until that moment, steps need to be taken with care and elegance to achieve the widest possible social benefit.

References

Barnett, H. (2002) *Constitutional and Administrative Law, 4th edition*, London: Cavendish Publishing.

Barraud, B. (2016) *La légistique. La recherche juridique*, L'Harmattan, https://amu.hal.science/hal-01367740/document.

Bingham, T. (2010) *The Rule of Law*, London: Penguin Publishing.

Brumfiel, G. (2023) "How AI is revolutionizing how governments conduct surveillance", [online], https://www.npr.org/2023/06/13/1181868277/how-ai-is-revolutionizing-how-governments-conduct-surveillance.

Buttice, C. (2023) "How AI Could Make Our Laws: First Bill Written by ChatGPT Gets Approved", [online], https://www.techopedia.com/how-ai-could-make-our-laws-as-the-first-law-written-by-chatgpt-gets-approved.

Campbell, J. (2008) *The Hero with a Thousand Faces*. New York: Pantheon Books.

Chevallier, J. (2015) "L'apport et le statut de la légistique", [online], https://hal.science/hal-01766821/document.

Chistou, P. (2023) " The Use of Artificial Intelligence (AI) in Qualitative Research for Theory Development", *The Qualitative Report*, DOI: 10.46743/2160-3715/2023.6536.

Enright, D (2019) "Politics Beyond the Ego. Ethical Political Leadership for the 21st Century", [online], https://www.routledge.com/rsc/downloads/CHBUSISP1901_LeaderShape_Politics_Beyond_the_Ego_WP_r3.pdf.

H. Judge P. Kevin Castel, Southern District of New York (2023) Order Case 22-cv-1461 (PKC), [online], https://s3.documentcloud.org/documents/23826753/judgeaskingtheotherlawyerwhyhesubmittedafilingwithfakecases.pdf.

Hill, G. (2023) "AI – Legislation for AI, Legislation with AI: Round No 1 ChatGPT v Australian Law Council", [online], https://ris.cdu.edu.au/ws/portalfiles/portal/94065165/Legislation_for_AI_Legislation_with_AI_ChatGPT_v_Australian_Law_Council_First_Round.pdf

Hutchins, J. (2003) "Machine translation and computer-based translation tools: what's available and how it's used", [online], https://www3.uji.es/~aferna/course4/MT%20process.pdf.

JURO (2023) The 2023 Tech GC report, [online], https://juro.com/library/2023-tech-gc-report.

JUSTICE (2023) "The State We're In: Addressing Threats & Challenges to the Rule of Law", [online], https://files.justice.org.uk/wp-content/uploads/2023/08/31123029/JUSTICE-The-State-Were-In-Addressing-Threats-Challenges-to-the-Rule-of-Law-September-2023.pdf.

Karpen, U. (2019) "Instructions for Law Drafting", [online], https://ial-online.org/wp-content/uploads/2019/11/Karpen-Instructions-for-Law-Drafting.pdf.

Keesey, R. (2018) *Modern parliamentary procedures, 2nd Edition*, American Psychological Association.

Marinese, V. (2008) "Légistique et effectivité", [online], https://doi.org/10.4000/books.pupo.1166.

McLachlan, S. (2024) "45 Facebook Statistics Marketers Need to Know in 2024", [online], https://blog.hootsuite.com/facebook-statistics/.

Millard, E. (2022) *Théorie générale du droit*, Paris: Dalloz.

Padjen, I. (2019) "Systematic Interpretation and the Re-systematization of Law: The Problem, Co-requisites, a Solution, Use", The International Journal for the Semiotics of Law, DOI: https://doi.org/10.1007/s11196-019-09672-x

Pszczyński, M. (2021), "Administrative Decisions in the Era of Artificial Intelligence", [online], DOI:10.14746/ppuam.2020.11.13.

Sanders, N. and Schneier, B. (2023) "How AI could write our laws", [online], https://www.technologyreview.com/2023/03/14/1069717/how-ai-could-write-our-laws/.

Stagnate Research (2023) " How AI is Revolutionizing Data Collection and Analysis", [online], https://stagnateresearch.com/blog/how-ai-is-revolutionizing-data-collection-and-analysis/

The Business Standard (2024) "Amazon advances automation with over 750,000 robots, replacing 100,000 jobs", [online], https://www.tbsnews.net/world/global-economy/amazon-advances-automation-over-750000-robots-replacing-100000-jobs-888656.

The UK Judicial Office (2023) "Artificial Intelligence (AI). Guidance for Judicial Office Holders" [online], https://www.judiciary.uk/wp-content/uploads/2023/12/AI-Judicial-Guidance.pdf.

Weinberg, A. (eds.) (2011) *The Psychology of Politicians*. Cambridge: Cambridge University Press

Westby, J. (2019) "'The Great Hack': Cambridge Analytica is just the tip of the iceberg", [online], https://www.amnesty.org/en/latest/news/2019/07/the-great-hack-facebook-cambridge-analytica/.

Artificial Intelligence and the Ethics of Tomorrow

Brenda van Wyk and Marlene Holmner
University of Pretoria, South Africa

Brenda.vanwyk@up.ac.za
marlene.holmner@up.ac.za

Abstract: Traversing our digital information society safely and responsibly rests mainly on our comprehension of the vast sociotechnical nature of AI ethics risks, its implications and consequences. Ultimately, we all would prefer to live in a mature information society that is technologically just, inclusive and sophisticated, firmly rooted in ethical information philosophy and values. In this paper the findings of a scoping review of recent reported research look, in particular, at the sociotechnical changes and impact that disruptive AI innovation has on societies, and how this could impact new and futuristic nuances in AI ethics. The study delves into the interdisciplinarity of AI ethics. The role of intergovernmental collaboration in researching and availing frameworks and guardrails in upholding AI ethics is critically interrogated and explored. The study alludes to gaps in current research around AI ethics and impresses the need to deliberate on future AI ethics dimensions. The prerequisites for fostering further confidence and trust in AI technology are synthesised. The study concluded that inclusivity and justice in AI Ethics is not yet achieved on a global level, and that there is still a tendency towards cultural and other biases in designing, planning, implementing and also regulating AI. More research is needed on the impact and trends of AI innovation in the Global South compared to the Global North.

Keywords: AI, AI Ethics, AI frameworks, AI interdisciplinarity

1. Introduction

The sociotechnical nature of AI and GenAI remains uncontested, as AI technologies are deeply entwined in societal factors, human behaviour, and ethical considerations. Navigating the current evolving AI environment requires technical skills as well as an awareness and understanding of AI ethics, the confounds of ethical risk and transgression consequences. The recent hype around AI, conversational AI and more specifically, generative AI (GenAI) may create a false impression that AI is the ultimate panacea for the future (Floridi 2020). On a more realistic note, Floridi (2019; 2020), as well as Sartori and Theodorou (2022) remind of the many challenges as well as breakthroughs experienced over nearly sixty decades of AI existence and innovations. The remarkable AI expansion and advancement of late, particularly in GenAI, is commonly referred to as an 'AI boom' or 'AI Spring' (Floridi, 2018, Sartori & Theodorou, 2022). The phenomenal advancement in large language models (LLMs) and prompt engineering are at the core of this hype, giving raise to new ethical concerns around cultural and language biases, but also on a personal and individual level. Finding answers for ethical dilemmas for the future should start with revisiting the philosophical roots of information ethics and AI ethics, and examining current trends.

1.1 AI and the Philosophies Underpinning AI Ethics Research

The study of AI ethics is considered part of applied ethics, and on a broader level, it is a branch of the philosophy of information. Floridi (2015) defines the philosophy of information as a field of philosophy specifically looking at critical research into the conceptual nature of information and its basic principles. Moral tenets and theories of ethics pertaining to all the aspects of AI can be complex, Revisiting the discourses of seminal authors affirms the foundational ethics knowledge, that could assist in addressing the ever increasing moral and ethical issues of AI.

1.2 Corner Stones in AI Development

The origins of the ideas on information ethics, which includes AI ethics, can be traced back to the early work of scientists and mathematicians, who interrogated ideas around computer ethics. One of these questions was: can computers be held responsible for their own acts, like humans are held responsible? In 1948 Norbert Wiener offered one of the first influential models in information ethics. His work laid the foundation for further discussions on the ethical implications of information and communication technologies. It addressed issues such as control, communication, and the responsibilities associated with the use of information in automated systems.

In early 1950, the computer scientists and futurists, Alan Turing he introduced the now infamous, 'Turing Test' to assess the intelligence of machines (computers). The Turing Test consisted of a human as judge who engages in a natural language conversation with both a human and a machine, without knowing which is which. He argued that the ability to engage in natural language conversation was a reasonable criterion for attributing

intelligence to a machine. The Turing test is once again under discussion today. What seemed like science fiction in 1950 is now our lived experience with the development of LLMs used in GenAI tools and applications such as ChatGPT. For the foreseeable future, values and norms remain humanistic and not yet exclusive technological constructs. Han (2022) reports that the oldest ethics and the ethical use of information in a digital environment are computer and information ethics, which in turn paved the way for new forms of ethics and new nuances in applied ethics. A seminal workshop was organised in 1956 by an interdisciplinary team of scientists in Dartmouth in USA, where the concept of artificial intelligence was coined (Sartori & Theodorou, 2021).

It is clear that the application of moral and ethical theories to AI technology can become intricate and complex. Earlier, Bibber's (1999) approach in addressing ethical concerns centred around negative normative constructs associated with technology, looking for what is undesirable, dehumanising, holds risks, or what deepens inequalities and creates exclusion and divisions. Recent views centre around what is desirable for the greater good of all within an information society.

Prominent frameworks developed from early research to inform new approaches in ethics, information ethics, computer ethics and ethics of technology research, teachings and practices. Just some of the noteworthy examples include the OECD's Fair Information Practice and principles (FIPPs) (1980), Floridi's Information Ethics frameworks, the Mason's (1986) PAPA Framework, and Tavani's Informational Privacy Framework (1999). Here, FIPPs are regulatory guidelines for safeguarding personal data. While Floridi's framework is more philosophical, and Mason's PAPA addresses specific ethical issues. Tavani's framework balances privacy with technological benefits. Though these examples are distinctly different and contribute their own unique focuses, they all add insights into the ethical handling of information, protecting privacy, ensuring accuracy, and promoting transparency and accountability.

Floridi (2016; 2018) alludes to the erroneous use of the term information society and presenting it as a single uniform place. In reality there is a distinction of several information societies existing on different levels of maturity with salient and variant human expectations and experiences. These rapid developments necessitate a more futuristic approach in deliberating the demands of the ethics of tomorrow. As such, the rationale of this review is to gauge the extent to which recent studies are offering evidence to inform what future ethics must guard.

2. Research Design

There has been a proliferation of published research on AI and AI ethics during the past six years. The purpose of this review is to gauge the coverage of recent peer-reviewed articles and conference proceedings to get a clear indication of the disciplines and their respective focuses on the influence of AI developments on ethics. The review considers interdisciplinary studies conducted between 2018 and 2023 on aspects of AI ethics. The question to be answered is: to what extent do existing frameworks and guidelines offered by international and intergovernmental organisations inform and guide AI ethics now and in the future? This study employs a scoping review for evidenced knowledge synthesis of articles with a strong philosophical underpinning.

2.1 The Search Strategy

Peer-reviewed articles in English, published between 2018 and 2023. Answering to the following criteria: (1) the context of the discussion is an AI-based or GenAI-based, (2) the discussion includes ethical considerations and how to implement or operationalise them into the AI-based application. Based on the recent escalation AI and GenAI developments (as reported by Floridi, 2019 and others) the study is limited to the past six years.

As an inter-disciplinary review, the disciplines with AI tangents and ethics were selected. The aim of this study is to consider a more generic and macro level look at the responsible use AI ethics in information societies. The search strategy covers philosophy, ethics, information science, business studies and computer science data bases. The databases included in the search strategy in this study are Google Scholar, Scopus and Web of Science. Databases were queried during June 2024. Databases were searched using keywords and search strings representing the two main concepts of AI technology and AI ethics. Websites were searched using the local search functionality and a combination of the following keywords: "Artificial Intelligence", "ethics", "Information ethics", "the philosophy of information ethics" and "Ethics of artificial intelligence" as systematically as possible within the search functionality and capabilities of each website. These results were then filtered according to document type and then further refined to look for aspects such as interdisciplinarity, applied philosophy of ethics, intergovernmental initiatives, inclusivity and the information society. A data log of the search strategies and results with links are archived in an Excel spreadsheet.

2.2 The Selection of Evidenced Sources

Guided by the PRISMA Extension for Scoping Reviews (Tricco et al. 2018), the flow chart in Figure 1 explains the knowledge synthesis of sources.

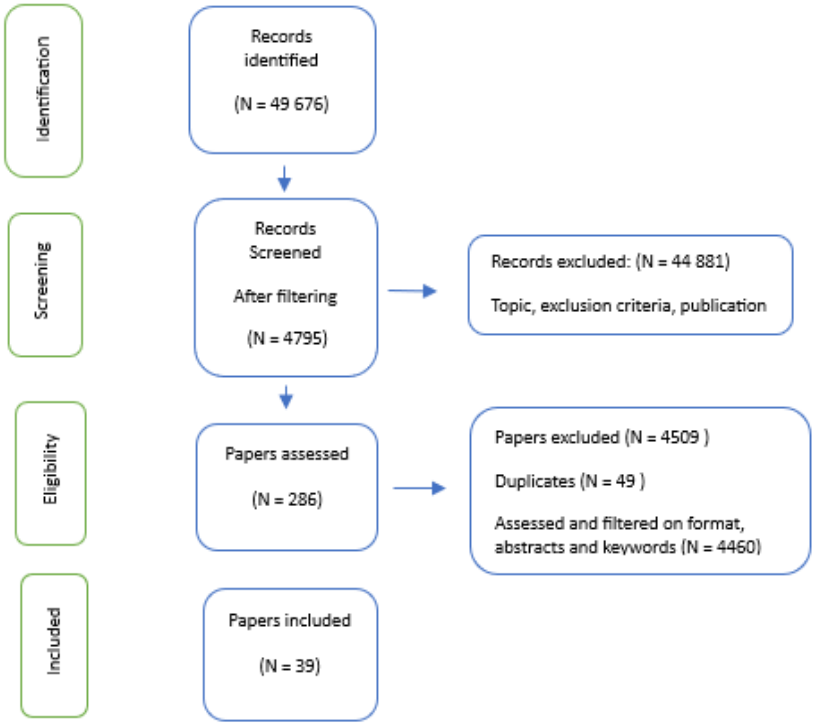


Figure 1: Scoping Review Process Flowchart

2.3 Inclusion and Exclusion Criteria

Inclusion criteria: Studies reported in full-text, peer-reviewed articles between 2018 and 2023, covering concepts of AI Ethics in the titles, abstracts or keywords were identified. Only studies in English were considered and selected.

Exclusion criteria: The following documents were excluded: grey literature, articles with abstracts only, patents, and reports, duplicate studies found in the databases.

2.4 Knowledge Synthesis of Eligible Sources

The first searches on Google Scholar, Web of Science and SCOPUS produced a result of 49 676 documents. After filtering, this was reduced to 4795 documents. In-depth screening for constructs of interdisciplinarity, philosophical and theoretical approaches, frameworks and guidelines used and reported, followed. With a total of 286 papers deemed to be eligible for this study, abstracts and keywords were screened for suitability of papers, resulting in 39 papers selected for the review. Reference links to included papers are only added to Table 1.

Table 1: Evidenced Sources included and used

Years	Types of research	Coverage	Reference
2019 -2022	Systematic literature reviews Literature reviews Typology research	A. Philosophy, Paradigms and Frameworks	A.1 Morley et al. (2020); A.2 Heiliner, (2022); A.3 Rességuier & Rodrigues (2021); A.4.Lauer, (2021); A.5 Hickok, (2021); I.2 Kazim & Koshiyama (2021);

Years	Types of research	Coverage	Reference
	Normative reflection		Also covered in: B.2 Hagendorff, 2020; C.3 Stenseke, 2022; C.4 Patel et al. (2019);
2018-2023	Systematic literature reviews Ethnographic research	B. Governance, guidelines and frameworks	B.1 Floridi (2018); B.2 Hagendorff, 2020; B.3. Butterworth, (2018) Also covered in: A.2 Morley et al. 2020; C.1 Bisconti et al (2023); F.2 Wright (2023)
2019 - 2023	Collaborative research	C. Interdisciplinarity, multidisciplinary	C.1 Bisconti et al (2023) ; C.2 Stenseke, 2022; C.3 Patel et al. (2019) ; C.4 Reeve, Zaugg & Zheng (2022)
2022 - 2023	Literature review	D. Education	D.1 Corrigan et al (2023); D.2 Raji, Scheuerman & Amironsei(2021) ; D.3 Bannister, Santamaría-Urbieta & Alcalde-Peñalver (2023); D.4 Holmes et al (2022); Also covered in C.5 Reeve, Zaugg & Zheng (2022); H.2 Chan (2023)
2019-2023	Systematic literature review	E. Sociotechnical nature of AI ethics	E.1 Sartori & Theodorou, (2022); E. 2 Floridi (2019); E.3 Green, 2021; E.4 Hagerty & Rubinov, (2019). Also covered in: G.1 Corrigan et al (2023); C.4 Patel et al. (2019);
2018 - 2013	Content analysis	F. Information Society, Mature information societies	F.1 Floridi et al (2018); F.2 Wright (2023); F.3 Bruneault & Sabourin (2021);F.4 Floridi (2020).
2023	Literature review Narrative inquiry	G. Justice and Inclusivity	G.1 Roche, Wall & Lewis, (2023); G.2 Hoffmann (2021); G.4 Park & Humphry (2019) Also covered in: D.2 Raji, Scheuerman & Amironsei, (2023); D.3 Bannister, Santamaría-Urbieta & Alcalde-Peñalver (2023);
2023	Bibliometric analysis	H. Generative AI Ethics	H.1 Bozkurt (2023) H.2 Chan (2023); H.3 Stahl & Eke (2023).
2021-2023	Systematic literature reviews	I. Human control, Human in the loop/ society in the loop	I. 1 Tasioulas; 2022; I.2 Kazim & Koshiyama; I.3 Akintunde (2023); I.4 Kyriakou & Otterbacher (2023); I.5 Riikka(2021); 1.6 Heyder et al (2023) Also covered in: C.1 Bisconti et al (2023)
2021	Bibliometrics	J. Digital identity	J.1 Zhang et al (2021); J.2 Cioroaica (2022), J.3 Bedushi (2019)

3. Discussion of Findings

The types of research studies within the final results included systematic literature reviews, scoping reviews, autoethnographic research, narrative inquiries, case studies, bibliometric studies and crowdsources research. Here, crowdsourced research is a model of the social organisation of research as a large-scale collaboration in which one or more research projects are conducted by multiple teams in an independent, yet coordinated manner. There has been a sharp increase in the number published studies on the topic of AI ethics in the Scopus database from 2021 to date.

3.1 Representative Countries

The results of this review indicated that most studies that met the criteria for inclusion, were conducted in countries within the Global North. These include the United Kingdom, USA, Canada, and Australia. This is in line with the extensive bibliometric study by Zhang et al (2021). They add that comparable research is being conducted in China and the European countries, but their key research communication is less accessible due to language barriers. There is a notable paucity in reported research emanating from Global South countries.

3.2 A. Philosophy, Paradigms and Frameworks in Researching AI Ethics

Many of the selected studies do not mention or elaborate on a particular paradigm or framework that informed their research. In their study, Roche et al (2021) profess an intersectional analytical framework to address the inherent complexities of AI ethics, and state that it is possible to understand how aspects of social and political identities combine to create different modes of discrimination and privilege.

3.3 B. Governance, Guidelines and Frameworks

Intergovernmental organisations, such as the OECD, UN and UNESCO offer valuable insights and guardrail on dimensions of AI ethics such as the inclusivity, education, awareness, towards further cooperation among countries and regions. In this review research (including Roche et al., 2022) report on the continuous efforts to address AI ethics risks, more particularly on social and cultural levels, resulting an ever-growing body of new knowledge that must translate into AI policies, frameworks and guidelines.

As early as 1980 the OECD's Privacy Guidelines outlined principles for the protection of privacy and personal data. Yet, today authors are critical and questions the intentions and contribution of intergovernmental organisations and their initiatives (Roche et al, 2021; Green, 2021; Floridi, 2018; Wright, 2023).

Wright (2023) is sceptical of organisations, governments and intergovernmental organisations who blatantly make use of “ethics-washing” practices in an effort to delay regulation of the sector. Green (2021) is in agreement, and warns of the gap between ethics as a mode of normative inquiry and ethics as a practical endeavour. There are suggestions of hiding political and normative disagreements. Rességuier and Rodrigues (2021) calls these guidelines “toothless” and ineffective. Conversely, Floridi (2018) argues that baseless and misplaced concerns may lead to underused AI technologies.

On the plus side, member countries have enacted legislation based on the information and guardrails offered, intended to protect individuals against the risk of AI abuse, and still give individuals the right of access to data with a view to checking their accuracy and appropriateness (Wright, 2023, 2021). The 2013 Privacy framework by OECD, and later in 2023 the OECD's Good Practice Principles for Data Ethics in the Public Sector serve as examples.

Hagerty and Rubinov (2019) underpin the need to ensure that emerging frameworks and standards themselves do not reproduce and reinforce a variety of biases, nor augment inequality. Authors (Hagerty & Rubinov, 2019; and Roche et al , 2021) allude to a lack of diversity as it is predominantly the research from Global North countries and research institutions are spearheading new developments in AI technologies and algorithms, bringing their own socio-cultural lenses. This homogenisation of applications in itself add to further exclusions and ethical biases.

3.4 C. The Interdisciplinarity, Multidisciplinary and Intersectionality of AI and GenAI Ethics

Patel et al (2019) state that by design, AI ethics must be interdisciplinary, as it considers broad pathways covering different disciplines to navigate AI innovation and ethics. Sartori and Theodorou (2022) add intersectionality as an important construct to the AI ethics debate. They contest that people from different backgrounds experience the world differently, be it due to gender, class, race or other forms of identity. This in turn affirms Floridi’s argument that we are dealing with a multiple of information societies. Stenseke (2022)

refers to these challenges as 'grand challenges' of global importance, impacting all aspects of human life, making the interdisciplinary approach critical.

3.5 D. Education

The centrality of education and training to the AI ethics debate delivered results within the search. Holmes et al (2022) remind that the core aim of education is to support learning with technology assisted teaching. AI concerns include fairness, accountability, transparency, bias, autonomy, agency, and inclusion (Holmes et al 2022). The compounded ethical challenges are present in the possibility of unintended consequences. It is important to distinguishing between differentiate between doing ethical things and doing things ethically in pedagogical choices.

3.6 E. The Sociotechnical Nature of AI and GenAI Ethics

Affirming the earlier studies of Bibber (1999) and Moore (2006), studies continuously report on the sociotechnical nature of AI that cannot be negated. The humanistic approach lies mainly in the sharing of values. Roche et al (2021) add that the rapid advancement of AI could have a positive and a transformative impact in individuals circumstances, as well as advancements in the broader economy.

3.7 F. Using AI Ethics Towards Maturing Information Societies

Florida (2019) criticises intergovernmental organisations' over-emphasis on technology. He laments that their narrow interest and remit focus on areas of development in broadband and telecommunications, the internet, consumer policy and digital governance. Floridi (2018) stresses that a mature information society centres around expectations. He states there is no global universal experience around digital innovation in the information society. We need to shift from viewing activities as 'off line' or 'online', to where we experience our world 'onlife' in an environment that Floridi (2018) calls the infosphere.

3.8 G. Justice and Inclusivity

The central ethics themes of AI include privacy and transparency. Hickock (2021) reports perspectives from different regions, and communities vary and the view that there are global and universal approaches to ethics is incorrect. Roche et al (2022) warn of ethnocentricity, whereby other cultures are evaluated through the standards and perspectives of the dominant culture. The Global South has to deal with historically unequal, colonial and biased social and power structures.

Reeve, Zaugg and Zheng (2022) warn that real-world harm is rife with biases such as racial profiling in online systems. Linking to education and interdisciplinarity of AI ethics teachings must urgently happen in computer science, data science, statistics, the social sciences and the humanities. Hagerty and Rubinov (2019) state that AI can deepen social divides, and exacerbating social inequality, particularly among historically-marginalized groups.

3.9 H. GenAI, Predictive AI and Conversational AI

Both GenAI and conversation AI , though distinctly different, are associated with AI risks. With rapid developments in GenAI, like ChatGPT, the need for human skills training is at the forefront. With prompt engineering attempting to make sense of human prompts, the concern for ethical problems increases. Immediate concerns are the risk to access and fairness. The AI and ethics of AI literacy communities should also focus on training individuals on the cognitive skills needed to become optimal users of such technologies. Bozkurt (2023) hails the educational potential of AI, but warns that it should be approached with caution to ensure responsible and ethical use.

3.10 I. Positioning Human Control in AI Autonomous Capabilities

The autonomous nature of AI is a huge concern for many, and there is a prevailing apprehension and fear for exclusion of human control. Kyriakou and Otterbacher (2023) emphasise the critical synthesis around the matter of human oversight. They refer to Rahwan who coined the human-in-the-loop (HITL) concept, and later society-in- the- loop (SITL) where human oversight can regulate AI and algorithm application and design as a form of social contract. Riikka (2021) alludes to inherent limitations and human error, and the inability to provide protection for fundamental rights. Akintunde (2019) states that replacing human decision-making with machine decision-making will be met with distrust in AI systems and moves to keep the human user in the loop.

3.11 J. Digital Identity, Ownership, Consent and Control

Beduschi (2019) warns that digital identity platforms may infringe human rights, and may infringe on privacy. It has the potential to be used in a discriminating manner. As digital identity refers to data informing our networked identity that uniquely describes and contains detailed information about a person, entity or business. Our digital footprint necessitates risk mitigation on both sociological as well as psychological levels. The particular AI threats of identity theft, deep fakes, to name but a few . Autonomy lies in user consent, ownership and control of data. Cioroaica et al (2022) posit that the need for ethical digital identities to fill the trust gap.

4. Conclusion

This review did a macro level exploration (see as explained Woodgate & Ajmeri, 2024) on the responsible use AI ethics, and also explore what is in stall for the near future. The development of AI has escalated significantly over the past six years. Research by scientists like Floridi and Satori, shows how deeply sociotechnical realities are interwoven in digital societies, and impress on us the need to think futuristically.

There are not many studies considering future AI ethics. Studies still concentrate on dealing with existing ethics challenges. The effectiveness of AI ethics to mitigate future harm lies in global efforts to sustain mature and informed information societies. Roche et al (2021) posit that AI ethics policies and standards must align with a variety of socio-cultural and socio-economic context. Regulation and accountability of AI ethics must continue to be visible and transparent.

The need for informed and clear AI ethics guidelines are bound to become increasingly complex, requiring foresight and expert knowledge of the technology, psychological and societal factors. AI as an emerging technology, has presented with various uncertainties, and the moral reaction on these developments are in many instances still fully or partially unknown. As we progress deeper into this era of technology, it is crucial to maintain a balance between innovation and accountability.

5. Recommendations

Based on the findings of the review, recommendations are made on how best to prepare to ensure the responsible use of AI in the future. The study recommends further research on how foresight analysis can be deployed in AI technology developments.

Floridi (2019) posits that predicting the ethics for the foreseeable future is often done with hindsight knowledge. There is a trend to build on fear and uncertainty:

> *"On the untestable side, some people will keep selling catastrophic forecasts, with dystopian scenarios taking place in some future that is sufficiently distant to ensure that the Jeremiahs will not be around to be proven wrong" (Floridi, 2019, 1)*

In a developing scenario, what the full extent of this maturity is, is not yet known. The infiltration of political and economic unethical opportunism in activities, finding their supple way into ethics frameworks and guidelines must have consequences. There are many unknowns, and Floridi (2019) states that it will not be innovation and change per se, but the regulation thereof that will determine the ethics of AI in the near future.

As AI advancements involve increased autonomous AI decisions, the singularity discourse must be on the radar of future ethics planning and frameworks. Floridi and Strait (2020) propose that ethics foresight analysis become more widely used. Ethical Foresight Analysis (EFA) is not a new concept in research and development, it refers to a variety of analytical methodologies for anticipating or predicting the ethical issues that technological artefacts may raise (Floridi & Strait, 2020).

References

Akintunde, M., Young, V., Yazdanpanah, V., Fathabadi, A.S., et al (2023). Verifiably Safe and Trusted Human-AI Systems: A Socio-technical Perspective. In Proceedings of the First International Symposium on Trustworthy Autonomous Systems (TAS '23). Association for Computing Machinery, New York, NY, USA, Article 56, 1–6. https://doi.org/10.1145/3597512.3599719

Bannister, P., Santamaría-Urbieta, A., & Alcalde-Peñalver, E. (2023). A Delphi Study on Generative Artificial Intelligence and English Medium Instruction Assessment: Implications for Social Justice. *Iranian Journal of Language Teaching Research*, *11*(3 (Special Issue)), 53-80. Doi: 10.30466/ijltr.2023.121406

Bindé, J., (2001) Toward an Ethics of the Future". *Globalization*, edited by Arjun Appadurai, New York, USA: Duke University Press, 2001, pp. 90-113. https://doi.org/10.1515/9780822383215-007

Beduschi, A. (2019). Digital identity: Contemporary challenges for data protection, privacy and non-discrimination rights. *Big Data & Society*, *6*(2). https://doi.org/10.1177/2053951719855091

Butterworth, M., (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework,

Computer Law & Security Review, 34 (2): 257-268, https://doi.org/10.1016/j.clsr.2018.01.004

Bisconti, P., Orsitto, D., Fedorczyk, F. *et al.* Maximizing team synergy in AI-related interdisciplinary groups: an interdisciplinary-by-design iterative methodology. *AI & Society,* **38**, 1443–1452. https://doi.org/10.1007/s00146-022-01518-8

Bozkurt, A. (2023). Unleashing the Potential of Generative AI, Conversational Agents and Chatbots in Educational Praxis: A Systematic Review and Bibliometric Analysis of GenAI in Education. Open Praxis, 15(4), pp. 261–270. DOI: https://doi.org/10.55982/openpraxis.15.4.609

Chan, C.K.Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *Int J Education Technology Higher Education* **20**, 38. https://doi.org/10.1186/s41239-023-00408-3

Cioroaica, E., Buhnova, B., Jacobi, F and Schneider,D., (2022). The Concept of Ethical Digital Identities, *IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI)*, Pittsburgh, PA, USA, 2022, pp. 17-20, DOI: 10.1145/3526073.3527586.

Corrigan, C.C et al. (Ed.) (2023) *AI Ethics in Higher Education: Insights from Africa and Beyond*, Springer International Publishing AG, 2023.

Floridi, L. (2013). *The philosophy of information*. Oxford University Press.

Floridi, L. (2015). *The ethics of information*. Oxford University Press.

Floridi, L. (2016). Mature Information Societies—a Matter of Expectations. *Philos. Technol.* **29**, 1–4 (2016). https://doi.org/10.1007/s13347-016-0214-6

Floridi, L. (2018). Soft Ethics and the Governance of the Digital. Philosophy & Technology, 31(1), 1–8. https://doi.org/10.1007/ s13347-018-0303-9

Floridi, L., Cowls, J., Beltrametti, M. *et al.* (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines* **28**, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L. (2019). What the Near Future of Artificial Intelligence Could Be. Philosophy & Technology, 32(1), 1–15. https://doi. org/10.1007/s13347-019-00345-y

Floridi, L. (2020) AI and Its New Winter: from Myths to Realities. *Philos. Technol.* **33**, 1–3. https://doi.org/10.1007/s13347-020-00396-6

Floridi, L., and Strait, A. (2020). Ethical Foresight Analysis: What it is and Why it is Needed? *Minds & Machines* **30**: 77–97. https://doi.org/10.1007/s11023-020-09521-y

Green, B (2021). The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice, in *Journal of Social Computing*. 2(3): 209-225, September 2021, Doi: 10.23919/JSC.2021.0018

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* **30**, 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8

Han, J. (2022). An Information Ethics Framework Based on ICT Platforms. *Information*, *13*, 440. https://doi.org/10.3390/info13090440

Heilinger, J.C. (2022). The Ethics of AI Ethics. A Constructive Critique. *Philos. Technol.* **35**, 61. https://doi.org/10.1007/s13347-022-00557-9

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI Ethics* **1**, 41–47. https://doi.org/10.1007/s43681-020-00008-1

Hoffmann, A. L. (2021). Terms of inclusion: Data, discourse, violence. *New Media & Society*, *23*(12), 3539-3556. https://doi.org/10.1177/1461444820958725

Holmes, W., Porayska-Pomsta, K., Holstein, K. *et al.* (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal Artifificial Intelligence Education,* **32**, 504–526. https://doi.org/10.1007/s40593-021-00239-1

Kazim, E and Koshiyama, A.S, (2021). A high-level overview of AI ethics, Patterns, 2 (9), https://doi.org/10.1016/j.patter.2021.100314

Lauer, D. (2021). You cannot have AI ethics without ethics. *AI Ethics* **1**, 21–25. https://doi.org/10.1007/s43681-020-00013-4

Mason, R. O. (1986). Four Ethical Issues of the Information Age. *MIS Quarterly*, *10*(1), 5–12. https://doi.org/10.2307/248873

Morley, J., Floridi, L., Kinsey, L. *et al.* (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science Engineering Ethics* **26**, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

OECD, (2013) The OECD Privacy Framework, https://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

Park, S., and Humphry, J. (2019). Exclusion by design: intersections of social, digital and data exclusion. *Information, Communication & Society*, *22*(7), 934–953. https://doi.org/10.1080/1369118X.2019.1606266

Patel, M., Webb., H, Jirotka, M., Davoust, A., Gales, R., Rovatsos, M and Koene, A. (2019). Harnessing Interdisciplinarity to Promote the Ethical Design of AI Systems. in P Griffiths & M Nowshade Kabir (eds),ECIAIR 2019 - Proceedings of European Conference on the Impact of Artificial Intelligence and Robotics. Academic Conferences and Publishing International (acpi), European Conference on the Impact of Artificial Intelligence and Robotics, Oxford, United Kingdom, 31/10/19.

Raji, I.D, Scheuermann, M.K., and Amironesei, R. (2021). You Can't Sit With Us: Exclusionary Pedagogy in AI Ethics Education, FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021: 515 – 525, https://doi.org/10.1145/3442188.3445914

Reeve, J., Zaugg, I. and Zheng, T. (2022), "Mapping data ethics curricula", *Journal of Information, Communication and Ethics in Society*, 20 (3): 388-399. https://doi.org/10.1108/JICES-12-2021-0124

Sartori, L., Theodorou, A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics Information Technology,* **24**, 4 (2022). https://doi.org/10.1007/s10676-022-09624-3

Stahl, B.B. and Damian Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology, *International Journal of Information Management*, Volume 74, https://doi.org/10.1016/j.ijinfomgt.2023.102700

Stenseke, J. (2022). Interdisciplinary Confusion and Resolution in the Context of Moral Machines. *Science Engineering Ethics* 28, 24. https://doi.org/10.1007/s11948-022-00378-1

Tasioulas, J. (2022). Artificial Intelligence, Humanistic Ethics. *Daedalus* 2022; 151 (2): 232–243. Doi: https://doi.org/10.1162/daed_a_01912

Tavani, H.T. (1999). Informational privacy, data mining, and the internet. *Ethics and Information Technology,* **1**(2), 137–145.

Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. (2018) PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. Doi: 10.7326/M18-0850

Woodgate, J., and Ajmeri, N. (2024). Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. ACM Computing Surveys, 1-39. Advance online publication

Wright, J. (2023). The Development of AI Ethics in Japan: Ethics-washing Society 5.0? *East Asian Science, Technology and Society: An International Journal*, 1–18. https://doi.org/10.1080/18752160.2023.2275987

Zhang, Y., Wu, M., Tian, G.Y., Zhang, G and Jie Lu, J, (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics, *Knowledge-Based Systems*, 222: 2021, https://doi.org/10.1016/j.knosys.2021.106994

Comparing Human-Labeled and AI-Labeled Speech Datasets for TTS

Johannes Wirth and René Peinl
Institute of Information Systems at Hof University, Germany

Johannes.wirth.3@iisys.de
Rene.peinl@iisys.de

Abstract: As the output quality of neural networks in the fields of automatic speech recognition (ASR) and text-to-speech (TTS) continues to improve, new opportunities are becoming available to train models in a weakly supervised fashion, thus minimizing the manual effort required to annotate new audio data for supervised training. While weak supervision has recently shown very promising results in the domain of ASR, speech synthesis has not yet been thoroughly investigated regarding this technique despite requiring the equivalent training dataset structure of aligned audio-transcript pairs. In this work, we compare the performance of TTS models trained using a well-curated and manually labeled training dataset to others trained on the same audio data with text labels generated using both grapheme- and phoneme-based ASR models. Phoneme-based approaches seem especially promising, since even for wrongly predicted phonemes, the resulting word is more likely to sound similar to the originally spoken word than for grapheme-based predictions. For evaluation and ranking, we generate synthesized audio outputs from all previously trained models using input texts sourced from a selection of speech recognition datasets covering a wide range of application domains. These synthesized outputs are subsequently fed into multiple state-of-the-art ASR models with their output text predictions being compared to the initial TTS model input texts. This comparison enables an objective assessment of the intelligibility of the audio outputs from all TTS models, by utilizing metrics like word error rate and character error rate. Our results not only show that models trained on data generated with weak supervision achieve comparable quality to models trained on manually labeled datasets, but can outperform the latter, even for small, well-curated speech datasets. These findings suggest that the future creation of labeled datasets for supervised training of TTS models may not require any manual annotation but can be fully automated.

Keywords: Text-to-Speech, Dataset generation, Pseudo labeling

1. Introduction

In the last few years, text-to-speech (TTS) technology has made remarkable progress, reaching previously unseen levels of naturalness and intelligibility. State-of-the-art (SOTA) systems consisting of neural networks have shown that they are capable of producing human-like speech that is almost indistinguishable from recordings of real human voices. Despite these remarkable results, the process of further developing and improving such systems becomes more demanding, especially when it comes to generation of new data.

The output quality of TTS systems strongly depends on large, high-quality datasets that are thoroughly labeled and well curated. However, the generation of such data represents a significant bottleneck, as it mostly involves manual labeling and quality assurance. For this reason, high-quality labeled speech data is scarce. This is especially true for non-English languages. While similar domains like automatic speech recognition (ASR) already benefit from techniques like pseudo labeling with ASR, TTS has not yet been equally considered regarding this method, despite speech data for TTS generally being of higher quality and thus also being easier to correctly be transcribed by ASR systems.

This work investigates the quality of pseudo-labeling TTS datasets using several speech recognition systems that involve both grapheme-based and phone-based approaches. The method described hereafter involves training SOTA TTS models on a dataset labeled by both humans and the previously mentioned ASR systems. We then conduct objective evaluations of intelligibility on model outputs using ASR. The results of these evaluations suggest that AI-labeled speech datasets are not only equally capable of labeling speech data with an accuracy comparable to human annotators, but in several cases can surpass the quality of human labeling. Our findings support the potential of pseudo-labeling as a powerful tool to improve the efficiency and effectiveness of TTS dataset generation.

2. Related Work

TTS systems have shown impressive progress since deep learning techniques have been applied to this challenge from 2016 on (Oord et al., 2016). The generation pipeline originally consisted of a Mel generator like the autoregressive Tacotron 2 (Shen et al., 2018) and TransformerTTS (Li et al., 2019), or the non-autoregressive like Glow-TTS (Kim et al., 2020) and Flow-TTS (Miao et al., 2020) and a vocoder to generate waveforms like Hifi-GAN (Kong et al., 2020) and WaveGrad (Chen et al., 2020). Later on, so called end-to-end TTS systems came into the focus that directly generate audio waveforms based on phoneme input like VITS (Kim et al., 2021) and JETS (Lim et al., 2022).

Most of these models were trained on the famous LjSpeech dataset (Ito, 2017), a single-speaker dataset with 24h of speech from the LibriVox speaker Linda Johnson. Single speaker data provided the best speech quality with a limited amount of training data. Multi-speaker datasets like VCTK (Yamagishi et al., 2019) which was recorded in studio-quality and TTS datasets curated from audiobooks such as M-AILABS (Solak, 2019), LibriTTS (Zen et al., 2019) and HiFi-TTS (Bakhturina et al., 2021) did not yield the same quality as single-speaker models although they contain far more data (586h in case of LibriTTS).

Only recently, large-scale TTS systems were developed, that are trained on thousands of hours of speech, that show clear advantages over single speaker models, especially in terms of prosody (Jiang et al., 2024). The required datasets are not human-labeled anymore but pseudo-labeled by ASR systems, e.g. LibriLight (Kahn et al., 2020), with its' nearly 60k hours of speech, which is used e.g. by VALL-E (Wang et al., 2023), MEGA-TTS 2 and NaturalSpeech 3 (Ju et al., 2024). The quality of the pseudo-labels is unknown, but due to low word error rates (WERs) of SOTA ASR models, especially for read speech with low noise, it seems good enough to produce high quality synthetic speech. However, this cannot be taken for granted, as the multi-lingual LibriSpeech dataset (MLS) (Pratap et al., 2020) for example has been shown to contain systematic errors in German language, e.g. for numbers and some text in the transcript, that was actually not spoken (Wirth and Peinl, 2022). In general, the quality of commercial and open source ASR systems have significantly advanced, besides English also in German (Wirth and Peinl, 2022). Although WERs can grow as high as 30% for challenging datasets like the multilingual TEDx Corpus (Salesky et al., 2021) (German subset), for read speech, good recording quality and low noise settings, you can expect WERs below 5% (e.g. M-AILBS: 4.28%, Voxforge: 3.36%, HUI: 1.89% with Conformer-Transducer large). For ASR systems, using pseudo labels has already been shown to provide benefits. (Hwang et al., 2022) demonstrated that a 600M parameter Conformer Transducer model delivers as good performance on ASR for voice search and conversations when being trained on over a million hours of pseudo labels in English language. (Zhang et al., 2024) use manually labeled data first (57k hours) and show that further training on up to 520k hours of pseudo-labeled speech reduces WER from 8.7% to 7.6%.

Despite the success and proven benefits of pseudo-labeling for the training of ASR models, this approach has not yet been adopted and evaluated in the field of TTS, which places even higher demands on audio quality and audio transcript alignment than ASR as well as phonemization of transcripts. Although transcript phonemization is not a necessity to train TTS models, phonemized texts enable more fine-grained control over the generation of synthesized speech outputs, as characters/graphemes can be pronounced differently depending on the context, while phonemes can be pronounced unambiguously.

In this work, we propose a phone-based ASR system to do pseudo-labeling and show, that it not only outperforms grapheme-based SOTA ASR systems (with algorithmic phonemization) for that task, but TTS models trained on the pseudo-labeled phones also produce better results than the one trained on the human-labeled reference data on multiple test sets.

3. Approach

To evaluate the quality of AI-based labeling for TTS datasets, a well-annotated and manually quality-assured speech dataset is selected to serve as a gold standard for subsequent comparisons. The audio recordings from this set are fed into a SOTA ASR model to create a variant of the original dataset with labels generated by the ASR model. Both the human-labeled and AI-labeled dataset variants are then phonemized and used to train two separate models with identical TTS architectures and hyperparameters, ensuring comparability.

Next, a test dataset consisting of text inputs is used to generate speech outputs from both TTS models. These speech outputs are then processed by another speech recognition system to produce transcript predictions. The predicted transcripts are compared to the original text inputs, and the WER is calculated for each model. By comparing the WERs, it is possible to determine how the performance of a TTS model trained on AI-labeled data compares to those trained on human-labeled data. While this type of evaluation does not substitute typical metrics for speech synthesis quality, such as mean opinion score (MOS), it effectively measures the intelligibility of synthesized speech output without the need for survey setups and participant recruitment (Peinl and Wirth, 2022). Although WER cannot be utilized to evaluate subjective speech factors such as the naturalness of the synthesized voice, the pace of speech or the pleasantness of the voice, as would be possible by means of a time-consuming evaluation of these factors with listener surveys and MOS-based ranking, it serves as a reasonable basis for determining the quality of synthesized speech in the scope of this work as general intelligibility of speech outputs is the decisive factor for the practical usability of a TTS system.

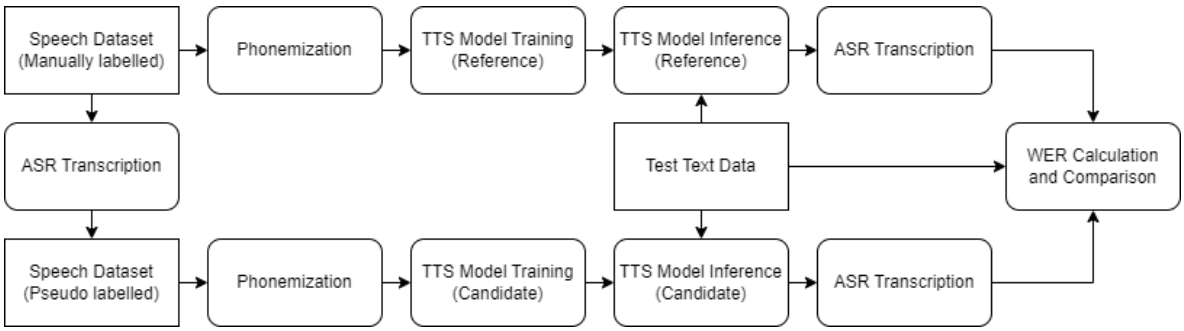


Figure 1: Flow Diagram of the approach used for comparison of human- and AI-labeled speech datasets for TTS systems. Entities are depicted in sharp corners, processes in round corners

As can be seen in Figure 1, the described approach is only illustrated for the comparison of a single pseudo-labeled dataset with an original dataset. The following sections present an implementation of this method with multiple pseudo-labeled datasets created using different ASR models.

4. Experimental Setup

The following subsections explain the specific measures established and models used to implement the previously described approach in detail.

4.1 Model Training

As reference dataset, the German single speaker dataset “Thorsten Voice” (Müller and Kreutz, 2022) was chosen, since it is of manageable size (~11hrs of audio-transcript pairs), has high audio quality and was labeled as well as quality-assured completely manually. In addition, multiple freely available TTS models have already been trained on this dataset, proving its high quality (Müller, 2024). Two different ASR models were used to pseudo-label the audio recordings of this dataset: The freely available Whisper v3 Large (Radford et al., 2022) model without utilizing the language detection mechanism (German was statically set as the transcription language) and an in-house model based on a SOTA Fast-Conformer Transducer Large architecture (Rekesh et al., 2023), which in contrast to Whisper uses a phone-based tokenizer and was trained on phone sequences from multiple languages (publication in preparation). This model does not include an explicit language detection mechanism but was trained with the aim of verbatim conversion of speech input into phoneme output, regardless of the input language, also enabling the model to create robust predictions with mixed language speech input. The phone representations are created using the international phonetic alphabet (IPA), whereas the grapheme-based Whisper model generates transcript predictions in plain readable text. The labels of the original dataset and the dataset pseudo-labeled using the Whisper model were further phonemized to IPA using a custom-built pipeline which, in addition to number and symbol normalization, also includes named entity recognition and language detection in order to ensure correct phonemization of loanwords from other languages. This is especially important for the German language, as it borrows many words and phrases from English and French. Punctuation was also removed from all transcripts to exclude it as a factor in the output quality of the models. The step of phonemization was skipped for the phone-based ASR model as it already creates outputs in the desired IPA format.

In the following step, each previously described dataset was used to train a VITS (Kim et al., 2021) model using the Espnet toolkit (Hayashi et al. 2020). Each TTS model was trained for 400 epochs, using an Nvidia RTX 6000 Ada graphics card.

4.2 Evaluation

Speech outputs were generated for each model to assess the quality of the TTS models pre-trained with the various dataset versions. These were then used as inputs for SOTA ASR models, and generated transcript predictions were compared with the original texts using WER to obtain an objective metric for the intelligibility of the TTS models.

Input texts were taken from the labels of the test splits of various speech datasets from diverse domains, including BAS HEMPEL (“BAS HEMPEL,” 2017), ASR Bundestag (Wirth and Peinl, 2023), Mozilla Commonvoice (Ardila et al., 2020), the Multilingual TEDx Corpus, M-AILABS, MLS, the Spoken Wikipedia Corpus (Baumann et al., 2019), the test split of the Thorsten Voice dataset used for training, Tuda-De (Radeck-Arneth et al., 2015), Voxforge (Voxforge.org, n.d.) and Voxpopuli (Wang et al., 2021). Since M-AILABS contains heterogeneous

speech data, on the one hand political speeches by former German chancellor Angela Merkel and on the other hand books read by single speakers, this dataset was further divided into the partial sets “M-AILABS” (excluding Merkel) and ”ASR Merkel“. For multilingual datasets only the German subsets were used.

Both Whisper V3 large, which was previously utilized to create a dataset variation with pseudo labels as well as a Conformer Transducer model pre-trained on German speech data provided by Nvidia (“STT De Conformer-Transducer Large,” 2022) were used to transcribe the synthesized speech outputs.

5. Results

The evaluation of TTS model outputs was carried out for every 100 training epoch checkpoints for all models; only the evaluation results for the epoch 400 checkpoints are listed below, as all models performed best at this stage, in comparison to scores of previous checkpoints.

Table 1: ASR evaluation results based on word error rate for the TTS models trained on the original dataset as well as pseudo labeled with Whisper v3 large and a phone-based ASR model (Whisper v3 large evaluation)

Dataset	Original	Phone Labelling	Whisper Labelling
HEMPEL	**11.82%**	11.95%	12.48%
ASR Bundestag	11.57%	**11.32%**	12.43%
ASR Merkel	6.76%	**6.37%**	7.51%
Commonvoice	18.63%	**17.40%**	20.82%
German TED	**12.54%**	14.54%	15.30%
M-AILABS	19.25%	**14.40%**	20.27%
MLS	14.26%	**13.46%**	15.90%
SWC	22.52%	**21.57%**	24.94%
Thorsten	**23.62%**	23.89%	25.78%
Tuda	11.63%	**11.15%**	14.29%
Voxforge	15.48%	**13.22%**	15.30%
Voxpopuli	**7.64%**	7.69%	9.45%
Mean	14.64%	**13.91%**	16.21%

Table 2: Evaluation results based on word error rate for the TTS models trained on the original dataset as well as pseudo labeled with Whisper v3 large and a phone-based ASR model (Conformer Transducer DE evaluation)

Dataset	Original	Phone Labelling	Whisper Labelling
HEMPEL	12.14%	**12.04%**	12.12%
ASR Bundestag	10.56%	**9.41%**	12.05%
ASR Merkel	5.76%	**5.74%**	6.36%
Commonvoice	15.30%	**14.29%**	16.79%
German TED	**12.57%**	14.32%	15.40%
M-AILABS	**11.47%**	11.56%	16.07%
MLS	9.92%	**9.72%**	11.93%
SWC	18.34%	**18.22%**	22.43%
Thorsten	**18.23%**	21.81%	24.42%
Tuda	**8.43%**	10.18%	12.58%
Voxforge	11.47%	**11.30%**	13.28%
Voxpopuli	**5.23%**	6.29%	7.48%
Mean	**11.62%**	12.07%	14.24%

As can be seen in Table 1 as well as Table 2, the TTS model trained on the pseudo-labeled dataset using a phone-based model exceeds the model trained on the dataset labeled by humans in most cases. Depending on the ASR model used for evaluation, these differences vary with the Whisper v3 results being slightly in favor of phone-based pseudo-labeling and the Conformer Transducer evaluation in favor of human labeling, which is

mainly attributed to the different datasets used to train these models. It can also be seen that all TTS models show comparatively small differences in terms of intelligibility in direct comparison, with the exception of few outliers.

While the model trained on original labels performed better on the test sets for Tedx, Thorsten Voice and Voxpopuli by small margins, it was outperformed in at least one model evaluation on all other test datasets. The model trained on pseudo-labeled data generated with Whisper v3 large did not achieve the best results on any test dataset, but its performance is not significantly worse across the board.

6. Interpretation

A direct comparison between TTS models trained on human- and AI-labeled data shows that all presented dataset labeling methods lead to very similar model performance levels. Notably, the model utilizing labels generated by the phone-based ASR system surpasses the one with original labels as well as the other variation trained on labels created with Whisper on most test datasets. This observation confirms the hypothesis that high-quality labeling of audio data with AI is possible and can even increase the quality of speech datasets based on the ASR model used to generate pseudo labels. We assume that the better model quality is not due to incorrect human labeling of the original dataset used, but due to the ability of the phone-based speech recognition model to represent verbatim spoken sounds better than is possible with grapheme-to-phoneme conversion. While the phonetic representation of most words is generated correctly by phonemization frameworks such as Phonemizer (Bernard and Titeux, 2021), it does not always correspond to what is actually spoken. Often, spoken words are shortened, concatenated or stressed differently depending on the context, which can lead to a greater discrepancy between actual spoken words in audio recordings and the phonetic transcript if a phonemization framework is used. This effect would be further reinforced by other factors such as spoken dialects. However, the potential cause of incorrect labeling described here can be avoided by using a phone- instead of a grapheme-based ASR model for pseudo-labeling and thus skipping the phonemization step in data preprocessing.

The high quality of the labels generated by the phone-based ASR model offers the additional benefit of its implicit multilingual pre-training. This eliminates the need to detect the language of the audio input or the necessity of maintaining multiple high-quality speech recognition models in different languages.

The difference in output quality between the two models trained on pseudo labels can be attributed to the rather logical than verbatim transcript outputs of the Whisper model in combination with the previously described induction of errors through phonemization. Since dataset labels used for speech synthesis should contain exactly what is actually said within a recording and not meaningful renditions, the phone-based model trained on verbatim output performs better on almost every test dataset in both evaluations. Conversely, this core finding may imply that phone-based speech recognition performs worse than grapheme-based speech recognition in the area of practically applied ASR, since in this area contextually correct transcription is in most cases more relevant than the exact transcription of what was said (including interfering factors such as repeated words due to stuttering and filler words), however, this was not investigated in the context of this work.

7. Conclusion and Future Work

In this work, it was shown that AI-based labeling of TTS datasets is not only a solid alternative to human annotation but is clearly the better option in the scope of the described approach and experiments. This underlines the effectiveness of modern ASR systems in generating high-quality labels that meet the requirements of TTS data preparation.

Based on these findings, the creation of new datasets could be fully automated, which can drastically reduce the development cycles of new models and increase the quality of speech synthesis systems in shorter iterations, enabling researchers to focus on model optimization instead of data quality assurance. In addition, since the best ASR model for pseudo labeling in the experiments conducted is phone-based and language-agnostic, the suggested and implemented approach is not only suitable for languages that have strongly pretrained ASR models but could also be used for data preparation in the scope of low-resource languages.

For future work, it is planned to further train the phone-based ASR model, which has achieved the best results for dataset labeling in the conducted experiments. By additionally improving the labeling quality, it will be possible to leverage large quantities of freely available, high-quality audio recordings. These recordings will be pseudo-labeled using the enhanced, phone-based ASR model and can subsequently be utilized to train a multi-

speaker state-of-the-art TTS model, specifically focusing on advancing the landscape of TTS for the German language and benefiting both researchers and end-users alike.

The phone-based ASR model including the custom built phonemization pipeline will be made freely available as soon as the accompanying research paper has been published in order make all presented results replicable.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common Voice: A Massively-Multilingual Speech Corpus, in: Proceedings of the 12th Language Resources and Evaluation Conference (Pp. 4218-4222).

Bakhturina, E., Lavrukhin, V., Ginsburg, B., Zhang, Y., 2021. Hi-Fi Multi-Speaker English TTS Dataset. arXiv preprint arXiv:2104.01497.

BAS HEMPEL, 2017.

Baumann, T., Köhn, A., Hennig, F., 2019. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. Language Resources and Evaluation 53, 303–329.

Bernard, M., Titeux, H., 2021. Phonemizer: Text to Phones Transcription for Multiple Languages in Python. Journal of Open Source Software 6, 3958. https://doi.org/10.21105/joss.03958

Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W., 2020. WaveGrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713.

Hwang, D., Sim, K.C., Huo, Z., Strohman, T., 2022. Pseudo Label Is Better Than Human Label.

Ito, K., 2017. The LJ Speech Dataset [WWW Document]. URL https://keithito.com/LJ-Speech-Dataset (accessed 5.3.21).

Jiang, Z., Liu, J., Ren, Y., He, J., Ye, Z., Ji, S., Yang, Q., Zhang, C., Wei, P., Wang, C., Yin, X., Ma, Z., Zhao, Z., 2024. Mega-TTS 2: Boosting Prompting Mechanisms for Zero-Shot Speech Synthesis. https://doi.org/10.48550/arXiv.2307.07218

Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.-Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., Zhao, S., 2024. NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models, in: ICML.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., Dupoux, E., 2020. Libri-Light: A Benchmark for ASR with Limited or No Supervision. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 7669–7673. https://doi.org/10.1109/ICASSP40776.2020.9052942

Kim, J., Kim, S., Kong, J., Yoon, S., 2020. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. arXiv preprint arXiv:2005.11129.

Kim, J., Kong, J., Son, J., 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech.

Kong, J., Kim, J., Bae, J., 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Advances in Neural Information Processing Systems 33.

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., 2019. Neural speech synthesis with transformer network, in: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6706–6713.

Lim, D., Jung, S., Kim, E., 2022. JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech. https://doi.org/10.48550/arXiv.2203.16852

Miao, C., Liang, S., Chen, M., Ma, J., Wang, S., Xiao, J., 2020. Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7209–7213.

Müller, T., 2024. Deutsche AI TTS-Stimme Kostenlos in Hoher Qualität ohne Cloud [WWW Document]. Thorsten Voice. URL https://www.thorsten-voice.de/kostenloses-deutsches-text-to-speech-tts/ (accessed 8.27.24).

Müller, T., Kreutz, D., 2022. ThorstenVoice Dataset 2022.10. https://doi.org/10.5281/zenodo.7265581

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Peinl, R., Wirth, J., 2022. Quality Assurance for Speech Synthesis with ASR, in: Intelligent Systems Conference (IntelliSys 2022). Amsterdam, Netherlands.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R., 2020. Mls: A large-scale multilingual dataset for speech research. arXiv preprint arXiv:2012.03411.

Radeck-Arneth, S., Milde, B., Lange, A., Gouvêa, E., Radomski, S., Mühlhäuser, M., Biemann, C., 2015. Open Source German Distant Speech Recognition: Corpus and Acoustic Model, in: Král, P., Matoušek, V. (Eds.), Text, Speech, and Dialogue, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 480–488. https://doi.org/10.1007/978-3-319-24033-6_54

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/ARXIV.2212.04356

Rekesh, D., Koluguri, N.R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., Hrinchuk, O., Puvvada, K., Kumar, A., Balam, J., Ginsburg, B., 2023. Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition. https://doi.org/10.48550/arXiv.2305.05084

Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D.W., Post, M., 2021. The multilingual tedx corpus for speech recognition and translation. arXiv preprint arXiv:2102.01757.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions, in: 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4779–4783.

Solak, I., 2019. The M-AILABS speech dataset [WWW Document]. URL https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/

STT De Conformer-Transducer Large, 2022.

Voxforge.org, n.d. Free Speech... Recognition (Linux, Windows and Mac) - voxforge.org.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., Wei, F., 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., Dupoux, E., 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. arXiv preprint arXiv:2101.00390.

Wirth, J., Peinl, R., 2023. ASR Bundestag: A Large-Scale political debate dataset in German. https://doi.org/10.48550/arXiv.2302.06008

Wirth, J., Peinl, R., 2022. ASR in German: A Detailed Error Analysis. arXiv preprint arXiv:2204.05617.

Yamagishi, J., Veaux, C., MacDonald, K., 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). https://doi.org/10.7488/ds/2645

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y., 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. arXiv:1904.02882 [cs, eess].

Zhang, K., Chkhetiani, L., Ramirez, F.M., Khare, Y., Vanzo, A., Liang, M., Martin, S.R., Oexle, G., Bousbib, R., Peyash, T., Nguyen, M., Pulliam, D., Donato, D., 2024. Conformer-1: Robust ASR via Large-Scale Semisupervised Bootstrapping.

PhD Research Papers

The Next Generation of AI Tools are Coming, do we Need to Better Regulate Them?

Michael Aubrey
University of South Wales, Blackwood, UK

04184025@students.southwales.ac.uk

Abstract: This research investigates the importance for robust regulation and legislation to properly govern the development of 'Frontier Artificial Intelligence Systems'. To do this a comparison of approaches of both existing and proposed legislation has been taken which includes from the US, UK and EU. This has involved reading both summations, assessments and opinions from both academic writers and the media and making comparisons with other industry regulations issues such as Social Media. The key findings of this research highlight the different approaches being taken to the same problem. Legislation that came into force on August the 1st 2024 from the EU takes a safety-first approach, identifying risk levels from 'unacceptable risk' that would be prohibited, down to 'minimal risk' which would remain unregulated. This is compared to the UK White Paper, which advocated an innovation first approach with a secondary focus on safety. With the potential risks associated with AI enhanced cyber-attacks and the spread of disinformation across different platforms, this research emphasises the importance of strong regulation in relation to safety and ensuring this happens from the outset in the development of 'Frontier AI' and is not an afterthought.

Keywords: Artificial intelligence, Cambridge analytica, Disinformation, Governance, Regulation, Safety

1. Introduction

The purpose of this research is to try to identify the current situation in relation to the governance and regulation of this industry in areas of global significance to the development of AI and compare their approaches. Then to investigate some of the potential issues that current AI technology creates and to speculate on how more powerful AI systems might accentuate these using examples that the currently unregulated social media market provides and to make some recommendations as to how governance and regulation may be improved based on the conclusions provided. It is a helpful to look at some of the lessons that we can maybe learn from the social media governance or lack of and to identify the issues and concerns in that area, as we can see integration from AI platforms such as 'Grok' into highly potent and influential Social Media platforms such as X. With social media being a relatively new phenomenon and a fast-evolving technology therefore using the experience with it will help us to gauge where we may see potential future issues from the progression of AI, its governance and regulation (Chauhan, Krishna Deo Singh. 2023).

There are numerous stories that exist in the global media extolling the virtues of Generative AI (Gen AI) and giving examples of its positive impact on, for example, education, where it can be used to improve efficiencies of staff and students in terms of planning and structuring lesson content and assignment work. This is of course a positive and using this type of technology for this type of purpose should be encouraged. However, we also see the negative where Gen AI is used to effectively cheat and claim credit for work that is not written, either by the student or by the academic but by the AI, and at some point, we are going to reach a point where either the positive outweighs the negative or vice versa, and what will this mean for trust, safety, data protection and the potential to spread misinformation and influence?

The expectations of this research are that it will likely show a significant area of concern around regulation especially when considering commercial interests. It is likely that significant lobbying has already taken place and will continue to do so, to attempt to ensure that there is a lighter touch to regulation in the development of these systems and ensuring the interests of technology companies is firmly represented. As frontier AI is likely to be a global phenomenon, as social media was back in the early part of the 21st century, it is likely that there will be significant global interest in harnessing the technology as quickly as possible. Frontier AI is described as an umbrella term for models of AI that outperform existing models in what they are capable of or task that they are able to perform (Jaquemard, 2023). This has been a long-standing goal for many developers of AI systems such as OpenAI, the company behind generative AI products such as ChatGPT. The current generation of AI tools already have the potential to cause huge problems throughout the world the question that we need to ask is will so called Frontier or General AI exasperate these problems? This is a complex question and in examining answers we need to first identify some of the potential issues that exist and how, through legislation and regulation, we might try to address these issues.

2. Review of Available Literature

Due to the ever-evolving nature of the technology behind AI and coupled with the fact this is a new field of study makes finding relevant academic sources in relation to the regulation of AI difficult especially as some of that regulation has not yet come to pass. However, we can use an existing technology like Social-Media (SM) and explore the approaches to that and how some of these lessons maybe learned in regulating AI development.

In order to see the effects of a perceived lack of regulation we only need to look as far as the SM market. In their article, Nyabola discusses at length the effect that social media has had on democracy and the challenges for regulating it. In its earliest conception SM was greeted with optimism and was seen as a great opportunity to build global connectivity, however has massive potential to undermine democracy in a number of ways for example the reinforcement of insular and damaging perspectives, appealing users inbuilt prejudices and the ability of bad actors through financial measures to buy greater visibility and influence and by exploiting a network that was developed with the best intentions (for example X) found a way to spread considerable misinformation and like we are now seeing with AI this creates a considerable regulatory challenge that as of yet has not been sufficiently addressed. Before the US election of 2016 the effect of SM in politics was seen as a complex issue, however after the election and the controversies that came with the influence SM had on the election of Donald Trump the challenge was seen as a global one rather that something that just affected less stable democracies (Nyabola, 2023).

We may ask, should we be concerned with the regulation of these industries, so to demonstrate the problem, we can point to examples such as the Cambridge Analytica (CA) scandal (Matara, M. et al. 2022). In 2016 two of the world's top five economic powers held significant democratic events in June the UK population were asked to vote in a referendum to decide the UK's membership of the European Union (EU) and in the US the presidential election was to take place in November. Ultimately these events produced results that had repercussions across the world, Brexit happened and of course Donald Trump was elected president. It later came to light that these outcomes may have been influenced by the collection of personal data through an app on Facebook called "my digital life". CA was ultimately able to gain access to data from it is estimated 87 million Facebook users. Even though the app only had an estimated 320,000 users, data was able to be harvested from not only these app users but also their friends and associates who they interacted with over Facebook, who crucially were unaware and had not consented to their data being used in this way. This allowed CA to create very accurate voter profiles and predict their voting intention this enabled them to better target individuals with advertisements that could influence an individual voter to vote the way CA or CA's clients wanted them too (Fadi Safieddine and Ibrahim, 2020), this was described effectively as a weaponisation of AI.

Shortly after the scandal broke the European Parliament made a resolution that recognised that the scandal had significant implications and the potential to manipulate past and future elections and referenda. In the UK, the House of Commons in 2019 issued a damaging report noting the influence of Russia on disinformation campaigns in 2016 that can be linked directly to the result of the Brexit referendum. The CA scandal bought to light fundamental weaknesses in existing legislation most notably GDPR regulation 679/216 specifies the right to protection of personal data, articles 7 and 8 of the Charter of Fundamental Rights and the EU-US privacy shield agreement of 2016 that sought to allow the easy transfer of data from the EU to certified companies in the US. After the scandal the privacy shield was no longer seen as a valid instrument for this purpose (Buruiană, Andreea, 2021). Whilst the weaknesses have certainly been highlighted adequate solutions have yet to present themselves, this will be partly down to the nature of legal challenges both to and from companies like 'Meta' and the fact that these companies were able to act with impunity with very little in the way of consequence highlights their global influence but also the significant weakness in legislation and regulation.

After the CA scandal it emerged that peoples trust in SM companies to keep their data secure was significantly compromised, Figure 1 shows we are seeing an erosion of trust in SM companies globally with only Governments being shown as less trustworthy as show in the Ipsos Global Trust monitoring report published in 2023. The figures in the report do indicate that trust is improving in all sectors however SM are still amongst the lowest scorers in this report.

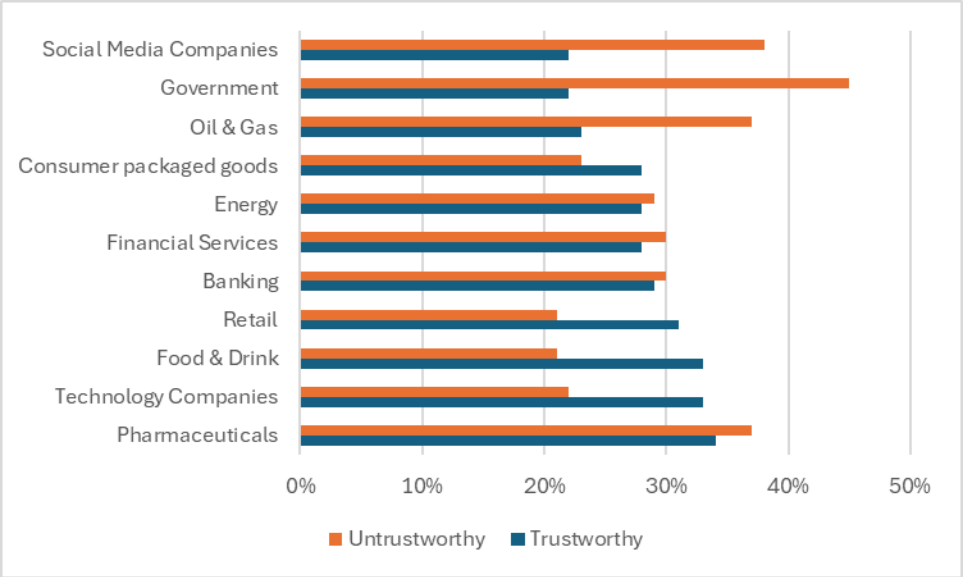


Figure 1: Global average levels of trust by sector in 2022/23 - (Ipsos, 2022)

What is interesting is the report also publishes the global attitude to regulation or lack of it, and once again SM companies score significantly high as shown in Figure 2 in terms of too little regulation.

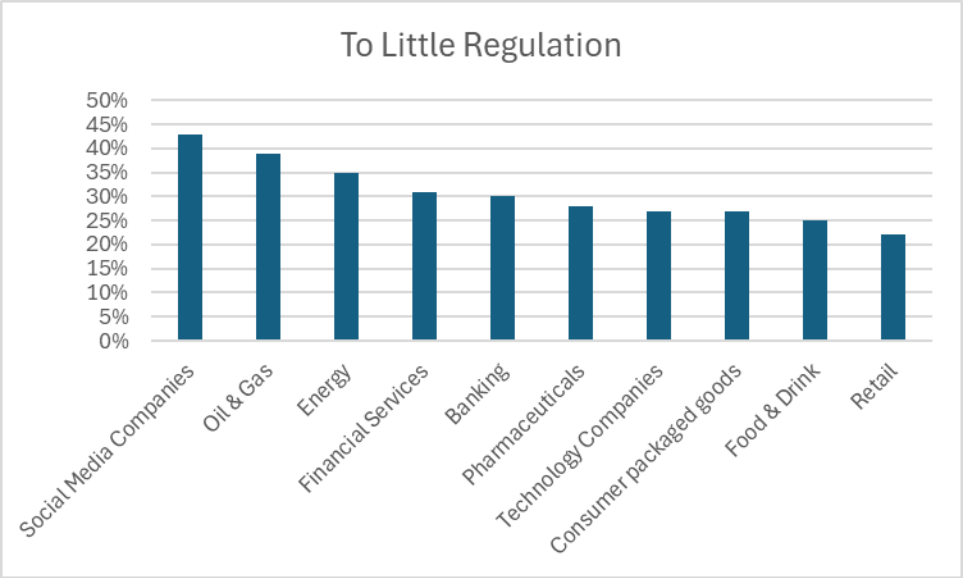


Figure 2: Global attitude toward regulation - (Ipsos, 2022)

These numbers are valuable for a number of reasons but in particular they show the value placed on trust and regulation. Although we would like more regulation, the irony is we do not seem to place trust in the people responsible for bringing about and enforcing that regulation i.e., Governments, as highlighted in Figure 1.

Unlike SM generative AI allows content that is no longer exclusively user generated but synthetic, and generated with great speed. Therefore, we can foresee many potential issues on the horizon where generative AI tools are used to created fake content or misinformation that is maybe not so damaging in isolation but when combined with the power of the social networks with multiple delivery channels this becomes a very different problem. In his article James Pinell makes this point and also espouses the need for parliaments around the world to act, also noting that;

> *This is as "bad" as these technologies are going to be. Artificial Intelligence is going to get more capable, not to mention more usable, over time. (Pinell, 2024).*

Whilst we may see Gen AI applications as powerful and advanced in their current form, they are only going to become more effective at what they do. This is a strong argument for governments to get a grip on this at the

earliest possible phase. In response to this the Commonwealth Parliamentary Association (CPA) have produced a handbook, designed to educate users on how to spot disinformation especially from synthetic generated sources that can be extremely difficult. This handbook contains strategies for combatting disinformation (Pinell, 2024), and whilst this is all a very welcome recognition of the problem and may go some way to help alleviate it, these actions can be simply ignored by bad actors who wish to spread this type of content.

3. Research Methodology

To substantiate the conclusions and recommendations made in this research a secondary methodology has been adopted using existing data and research from credible journalistic and academic sources. The scope of the paper brings in sources within a four-year time frame from 2020 although there is reference to work done from as far back as 2016. Geographically the focus is primarily on the United Kingdom (UK) and Europe however where pertinent references are made to China, the United States (US) and Australia. In searching various academic databases, although some work has been done in this area there is not a large amount of relevant research currently in existence, this is put down to the recent emergence of the technologies under discussion and it is assumed that more published sources will become available in the future.

4. How is AI Currently Regulated or Proposed to be Regulated?

Ultimately it is not the job of legislators to restrict and stymie technological breakthroughs, but it is their job to ensure these are done safely, ethically and responsibly (Jacobs, M. & Simon, J. 2023). In his article entitled *'AI Regulations Around the World: A Comprehensive Guide to Governing Artificial Intelligence'*, Savio Jacob highlights the different approaches currently being taken by different governments around the world. For example, the UK has produced a White Paper that proposes an innovation led approach to governance and regulation, whereas, US is currently taking a decentralised approach and has several sector specific agencies that set out to address the challenges brought on by AI development. In the EU, regulation is broken down in risk factors and very much geared to placing accountability for responsible development on the companies who develop these models first and the people who use the technology second. China is approaching this differently in the wish to remain at the forefront of AI technological developments, indeed their focus is on maintaining their superiority in AI and the plan under the Chinese Cybersecurity Law and the New Generation AI Development Plan integrated both security and data protection measures. Australia has produced the 'National Artificial Intelligence Ethics Framework', which direct the ethical principles by which AI technologies are developed and implemented, this approach focuses away from the data protection element for the individual user and places more emphasis on how the technology is developed (Jacob, 2024).

In March 2023 the Government of the UK published its White Paper entitled 'A Pro Innovation Approach to AI Regulation", indeed White Papers are often seen as a precursor to specific policy directions by which the government of the day wishes to pursue. The title of this specific paper championing the phrase 'Pro Innovation' is potentially problematic when it comes to developing a safety-first approach to regulation. Where innovation should certainly not be frustrated by regulation, striking a balance between what is innovative and what is ethical should always be considered, for example it could be argued that the approach CA took in 2016 to its campaigns was indeed innovative but was certainly not ethical. Therefore, we must throw a note of caution in pursuing an innovation led approach to regulating this industry, not simply allowing progression before safeguards are put in place to ensure everyone's safety and to this transparency is key.

In the White Paper the UK government recognises the importance of AI and the benefits already being seen in society, and it identified AI as being one of the five critical technologies. It recognises that across the world government are beginning to legislate for this and Sir Patrick Vallance (the then UK Governments Chief Scientific Advisor), notes that this is very much a time critical venture if the UK is to take full commercial advantage in a pro-innovation approach to make the UK an attractive place to encourage AI start-ups. The Paper emphasises the risks to both physical and mental health in addition to the need for data protection and privacy as well as the protection of human rights. Building public trust is also seen as a key factor in this White Paper, public trust is seen as a key to effective regulation with the need for clear and consistent support where government investment i.e. taxpayers' money is used. However, it is important to accelerate the adoption of AI across the UK, again putting a pro-innovation strategy at the forefront and promoting AI technologies in the most positive light possible. The danger in this approach that significant risks are either overlooked or simply dismissed. There is recognition that currently AI as a general-purpose technology is regulated across a complex framework of requirements in different sectors and industries and this is a concern and therefore the paper recognises the need for a clear and unified approach to regulation (UK Government, 2023).

In the US, the approach to governance is far more fragmented due to the nature of having both Federal and State Laws. It is entirely possible without blanket legislation for individual states to pass their own AI bills meaning that different states in the US will regulate in different ways. Currently in the state of California which is one of the earlier adopters of generative AI and with plans to help use it for things such as traffic management and highway monitoring, Senator Scott Wiener has authored a Bill proposing to create a stronger regulatory requirement, setting out further safety standards that would be designed to ensure models are thoroughly tested and to bring forward measures to prevent large AI models from being used to create potential catastrophes such as manipulation of the power grid or the building of chemical weapons (Nguyen, 2024).

On the 1st August 2024, the AI Act came into force across the EU. It classifies AI according to its assumed risk and identifies four categories of risk by which the Act seeks to impose regulation:

- **Unacceptable risk**
- **High risk systems**
- **Limited risk systems**
- **Minimal risk systems**

This Act pushes out further than just the boundaries of the EU by dictating that products intended to be put into service inside the EU regardless as to where they are based, are obliged to conform to the regulations laid down in the Act (Stewart, L. S. 2024). This also includes the products outputs, so even output generated from a system based outside of the EU, if that output is to be used within the EU it is covered by the Act (European Union, 2024).

2.1 Unacceptable Risk

Under the Act these are some examples of types of AI systems are prohibited i.e. not permitted and are covered by Article 5 of the Act. Unacceptable risk is likely to be the most challenging category to enforce unless the categorisation is very clearly defined, under EU law this would prohibit the development of systems that fall into this category and may bring legal challenges from any company whose systems are categorised in this way.

2.2 High Risk

Under Annex III of the AI Act these are some examples of areas that AI systems could be developed and deployed that are deemed high risk and therefore subject to significant regulation. The areas covered under the high-risk classification include some sectors under government control. Protecting critical national infrastructure such as transportation, power and water supply would be seen as a priority and therefore the introduction of AI systems into the running of any of these would need to be independently monitored. This would be especially true in terms of power generation where private investment and influence is substantial and where the need to generate profit is seen as a priority.

2.3 Penalties

Unlike so many Acts that are put in place to regulate and govern industries this one does have some teeth and is able to levy significant penalties where appropriate. For example, practices relating to Article 5 i.e., the prohibited Annex of the Act can lead to fines being levied in the region of €35,000,000 or, should the offender be proposing or promising something that is deemed prohibited 7% of previous annual worldwide turnover (whichever is higher). Where non-compliance is deemed according to elements of the Act not covered in Article 5 then fines of up to €15,000,000 or, if there is an undertaking then 3% of previous annual worldwide turnover (whichever is higher).

2.4 Codes of Practice and Governance

The AI Act does not introduce a Code of Practice but does make provision for one to be drawn up under article 95. It suggests that the AI office and member states collaborate to develop this as a voluntary application to governance of AI systems that are not deemed high risk (these systems are already covered under Chapter III of the AI Act). Article 95 also suggests that individual codes of practice may be drawn up by providers, deployers of AI systems or by interested stakeholders or representative organisations. This gives cause for concern as to how effective these could be, although it is recognised that producing a standardised one size fits all Code of Conduct may prove restrictive and prohibitive enough to encourage non-compliance or take away the incentive to sign up altogether.

2.5 ISO 42001

The official title of this standard is ISO/IEC 42001, Information Technology – Artificial Intelligence – Management System. The standard seeks to specify the requirements and to provide guidance to companies in relation to implementing AI systems for management purposes within their organisations and was published in 2023. This Standard does not set out best practice for the development of AI but more specifically how it is implemented and managed in the managements systems of an organisation. This is an important, although not fool proof, safeguard and with the growth in the use of AI by large companies and governments having an agreed standard to work to, will provide some semblance of oversight.

The Standard intends to help organisations to be responsible in their approach to the implementation of AI with specific considerations given to:

- The use of AI for decision making
- The use of AI for data analysis
- AI systems that perform continuous learning and change their behaviours during use

The Standard gives expectations to organisations to focus their effort into these areas to try to always ensure transparency in relation to how AI is being used and to attempt to mitigate the risks involved (Internation Standards Organisaton, 2023)

2.6 The 1st AI Safety Summit, November 2023

The AI Safety Summit was convened by the then Prime Minister of the UK Rishi Sunak in November 2023 at Bletchley Park, a site made famous after World War 2 as being the place where many Axis Codes were broken and certainly known to be one of the first uses of what we might recognise as modern computer systems. The Summit was attended by representatives from 29 different governments including the EU and the United States. The overarching aim of the Summit was to attempt to garner a joined-up approach in the future developments of AI, especially the development of “Frontier AI”.

2.7 The Bletchley Declaration

With Summits of this type, it is highly unlikely that we will get substantial and robust policy made, that all participants can agree on and where real change can happen. This Summit appears to have been no exception where a list of “nice to haves” was produced but not backed up by any enforcement, regulation or guidance to companies who are developing these advanced models. The Summit recognised, quite correctly the use of AI to enhance our lives whilst also recognised the need for safe design and how the next generation of frontier AI should be very ‘Human Centric’ (Government of the United Kingdom, 2023). It also recognises there are unseen risks. Referring to its concerns over current AI systems to manipulate or generate content that is inaccurate or deceptive in nature and notes the urgency of addressing these issues. What the Summit does not do in any substantial way is plot a route to any potential governance of this industry, indeed, governance and regulation is not mentioned until the third page of the document. When this is mentioned, it is vague, discussing applicable legal frameworks and national circumstances and seems to only pay lip service to the potential problem.

2.8 The 2nd AI Safety Summit, May 2024

The 2nd AI Safety Summit was held in Seoul, South Korea, was co-hosted by the Republic of Korea and the government of the UK. The outcome of the Summit was criticised by some in diluting the purpose of the Summit by adding additional topics for discussion other than simply AI safety, including innovation and inclusivity. Whilst this Summit could be seen as largely irrelevant, especially when compared to its previous outing it did produce several significant points of discussion. Most notably on AI safety institutes and their increasing number giving increased capacity across governments to better understand and regulate at both national and international levels. The addition of further AI safety institutes across countries such as South Korea, Japan and Canada when added to the already declared institutes in the UK and US, the European Commission AI Office serves as the AI safety institute for its 27 member countries, the Summit was able to secure the backing of 10 countries plus the EU for greater cooperation. This marks the true legacy of this Summit as to realising international cooperative efforts in relation to AI safety (Allen and Adamson, 2024).

2.9 AI companies and Their Responsibilities - Safety Committees

As noted in the EU AI Act responsibility is placed on the developers of this technology to develop systems that is not only safe but ethical. If we look at some of the most significant companies in the field, for example,

OpenAI in May 2024 introduced its Safety and Security Committee. This Committee is primarily responsible for reporting to the full board on critical safety issues and security discussion. This Committee is also tasked with reviewing the processes and safeguards around the development of new frontier AI models, reports from this Committee are set to be published in line with their recommendations to the board of OpenAI (OpenAI, 2024).

The Safety Committee for OpenAI, should ensure that the company does not go beyond what is deemed acceptable by law, however, there is a significant weakness in the makeup of the Committee. If we examine the members of the Committee it is comprised of existing board members which includes Sam Altman the CEO of OpenAI. Whilst it is not impossible for this Committee to do its job in a transparent manner, it must also be pointed out that there are conflicts of interest here and the lack of a truly independent overseer is significant.

3. Conclusions and Recommendations

Whilst researching this topic it became clear that that is a significant lack of a cooperative approach across the world to regulating anything, what is in one countries interest is not in another's, however, when it comes to technology such as AI, which has such potential, seeing this technology being exploited for any nefarious purpose is troubling. The question that has been running through this Paper is, of course, what we as a global community do about it. If we look at other areas of technology that have such potential for good, but also harm, and see how this is done and we may start to be able to develop answers to this problem. Atomic energy for example, has an international regulator, the International Atomic Energy Agency (IAEA) that has been on statute since 1956 with the United Nations and has a briefing to promote best practice, safety standards and to assist all economies in applying these standards. It would seem, a similar approach could be taken with AI development and therefore based upon this research it is recommended that:

- There is an implementation of an International Agency responsible for the application of agreed standards in AI development, this Agency needs to be properly funded through a similar model to the IAEA and have powers of inspection and sanction
- The Agency should have the power to launch a surprise inspection, providing it can be shown it has clear grounds to do so
- The Agency should seek to implement a Standardisation Programme that could be ISO42001, or based upon this, that AI developers should be able to sign up for and be regularly assessed to ensure that they are meeting these set standards
- Standards should be reviewed over a regular period, which is to be defined, to ensure that the Agency is working with the most up to date information and models

Whilst setting up an Agency of this type may prove significant, like all codes of practice it will require buy in from everyone to make it truly successful. However, the phrase ‘do not let the perfect be the enemy of the good’, is applicable if in doing this we can considerably reduce the risks posed by rogue AI development.

Other areas of the world that are likely to have significant influence on the development of AI include both the EU and the UK. However, the approach to AI regulation is significantly different between the two, the White Paper published by the UK seems to prioritise innovation above safety, which implies a potentially lighter regulatory touch. This approach is likely designed to encourage companies to locate the research and development arms in the UK as opposed to other nations thus ensuring significant economic benefit for the UK. Whereas the EU approach is seen as significantly more safety oriented, better defining both risks and penalties and the creation of an overseer that will have power at Federal level to impose significant penalties.

We have already seen the potential damage that even the less sophisticated AI type systems can do or at least have the potential to do in the CA scandal that came to light in 2016. Although the argument is that this was not true AI, the building blocks and potential can be easily observed in the use of algorithms designed to specially characterise individuals based on learned criteria. What was also apparent from this scandal is that although personal data is protected under frameworks such as GDPR this was not a significant deterrent and was exploited significant (Delcker, J. 2020).

Acknowledgements

This work was supervised and supported by Dr Richard Ward and Dr Mabrouka Abuhmida of the University of South Wales and Mrs Dawn Bradley of Coleg Gwent for proofreading.

References

Adamson, G. C. Allen, G. (2024). *The AI Seoul Summit*. Retrieved July 22, 2024, from https://www.csis.org/analysis/ai-seoul-summit

Buruiană, A. (2021) THE LEGAL CONSEQUENCES OF THE FACEBOOK-CAMBRIDGE ANALYTICA SCANDAL ON THE EU INFORMATION SOCIETY; THE PROPOSAL OF E-PRIVACY REGULATION. *Analele Universității Titu Maiorescu. Drept*. XX (XX), 47–58.

Delcker, J. (2020). *AI: Decoded: How Cambridge Analytica used AI*. Retrieved July 22, 2024, from https://www.politico.eu/newsletter/ai-decoded/politico-ai-decoded-how-cambridge-analytica-used-ai-no-google-didnt-call-for-a-ban-on-face-recognition-restricting-ai exports/#:~:text=HOW%20CAMBRIDGE%20ANALYTICA%20USED%20AI%3A%20Artificial%20intelligence%20p

Chauhan, K. D. S. (2023) From 'What' and 'Why' to 'How': An Imperative Driven Approach to Mechanics of AI Regulation. *Global jurist*. [Online] 23 (2), 99–124.

European Union. (2024). *AI Act Explorer*. Retrieved July 16, 2024, from https://artificialintelligenceact.eu/ai-act-explorer/

Government of the United Kingdom. (2023). *About the AI Safety Summit 2023*. Retrieved May 29, 2024, from https://www.gov.uk/government/topical-events/ai-safety-summit-2023/about

Ibrahim, Y. & Safieddine, F. (eds.) (2020) *Fake news in an era of social media : tracking viral contagion*. London ; Rowman & Littlefield International.

International Standards Organisaton. (2023). *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system*. Retrieved August 22, 2024, from https://www.iso.org/standard/81230.html

Ipsos. (2022). *IPSOS GLOBAL TRUSTWORTHINESS MONITOR, STABILITY IN AN UNSTABLE WORLD.* Ipsos.

Jacob, S. (2024). *AI Regulations Around the World: A Comprehensive Guide To Governing Artificial Intelligence*. Retrieved July 10, 2024, from https://www.spiceworks.com/tech/artificial-intelligence/articles/ai-regulations-around-the-world/

Jaquemard, T. (2023). *Frontier AI: Heading safely into new territory*. Retrieved May 29, 2024, from https://trilateralresearch.com/emerging-technology/frontier-ai-heading-safely-into-new-territory

Jacobs, M. & Simon, J. (2023) Reexamining computer ethics in light of AI systems and AI regulation. *Ai and ethics (Online)*. [Online] 3 (4), 1203–1213.

Matara, M. et al. (2022) *Facebook and Cambridge Analytica : an unholy alliance*. London: The Eugene D. Fanning Center for Business Communication, Mendoza College of Business, University of Notre Dame.

Nguyen, T. (2024). *California advances unique safety regulations for AI companies despite tech firm opposition*. Retrieved August 5, 2024, from https://apnews.com/article/california-meta-google-ai-regulation-fight-e3a9c25b597b5f469fe0de149572f87e

Nyabola, N. (2023) Seeing the forest – and the trees: The global challenge of regulating social media for democracy. *The South African journal of international affairs*. [Online] 30 (3), 455–471.

Open AI. (2024). *OpenAI Board Forms Safety and Security Committee*. Retrieved July 22, 2024, from https://openai.com/index/openai-board-forms-safety-and-security-committee/

Pinnell, J. (2024) AI, DISINFORMATION AND DEMOCRACY: THE NEED FOR PARLIAMENTS TO ACT. *Parliamentarian*. 105 (1), 36-.

Schneble, C. O. et al. (2018) The Cambridge Analytica affair and Internet-mediated research. *EMBO reports*. [Online] 19 (8), .

Stewart, L. S. (2024) The regulation of AI-based migration technologies under the EU AI Act: (Still) operating in the shadows? *European law journal : review of European law in context*. [Online] 30 (1/2), 122–135.

UK Government. (2023). *A pro-innovation approach to AI regulation*. Retrieved July 22, 2024, from https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#executive-summary

“Should Everyone Have Access to AI? " Perspectives on Ownership of AI Tools for Security

Yasmine Ezzeddine and Petra Saskia Bayerl
Sheffield Hallam University, Sheffield, UK

y.ezzeddine@shu.ac.uk
p.s.bayerl@shu.ac.uk

Abstract: Given the widespread concerns about the integration of Artificial Intelligence (AI) tools into security and law enforcement, it is natural for digital governance to strive for greater inclusivity in both practice and design (Chohan and Hu, 2020). This inclusivity can manifest in several ways, such as advocating for legal frameworks and algorithmic governance (Schuilenburg and Peeters, 2020), allowing individuals choice, and addressing unintended consequences in extensive data management (Peeters and Widlak, 2018). An under-reflected aspect is the question of ownership, i.e., who should be able to possess and deploy AI tools for law enforcement purposes. Our interview findings from 111 participants across seven countries identified five citizens viewpoints with respect to AI ownership of security-related AI: (1) Police and police-governed agencies; (2) Citizens who disassociate themselves; (3) Entities other than the police; (4) All citizens including themselves; and (5) No one or Unsure. The five clusters represent disparate perspectives on who should be responsible for AI technologies, as well as related concerns about data ownership and expertise, and thus link into broader discussions on responsibility for security, i.e., what deserves protection, how and by whom. The findings contribute theoretically to digitalization, smart technology, social inclusion, and security studies. Additionally, it seeks to influence policy by advocating for AI development that addresses citizen concerns, thereby mitigating risks, social, and ethical implications associated with AI. Crucially, it aims to highlight citizens’ concerns around the potential for malicious actors to exploit ownership of such powerful technology for harmful purposes.

Keywords: Artificial intelligence, Ownership, Citizens, Law enforcement agencies, Police

1. Introduction

In the domain of security and policing, the integration of Artificial Intelligence (AI) tools presents both unprecedented opportunities and ethical considerations. At the crux of these advancements lies a pivotal question: who should hold ownership of these AI tools, and thus who owns the responsibility for security?

Expanding the definition of AI is crucial, considering recent entries defining AI as "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals" (European Commission, 2020). With AI tools rapidly expanding across society, including in the security and policing domain, digital governments are seeking more inclusive dynamics in practice and design to ally valid citizen concerns (Chohan and Hu, 2020). This inclusivity can manifest in various ways, from calls for dedicated legal frameworks to algorithmic governance and better accounting for unintended consequences in large data management (Schuilenburg and Peeters, 2020; Peeters and Widlak, 2018).

An underexplored aspect in this regard is AI ownership. Generally, AI in the security and policing domain is conceptualised as police-owned capability. However, the ongoing privatisation and personalisation of security (for instance privately owned door cameras with AI capabilities) opens the field to a much more fluid landscape of ownership. This study explores the perceptions and preferences of citizens regarding the ownership of AI tools, particularly in policing and law enforcement contexts. Understanding public stances on this matter is crucial, as they reflect societal values on societal understandings about responsibilities for security (Bayerl et al., 2022) and can therefore contribute to shaping policies governing AI deployment and ownership.

In the realm of AI and societal implications, the discourse on ownership of security-related AI capabilities has been a focal point in academic and policy discussions. Numerous studies have explored AI deployment landscapes. However, these discussions often overlook citizen perspectives regarding ownership of security-related AI. While scholars and policymakers deliberate on governance models and ethical frameworks, citizen voices remain underrepresented in these discussions (Floridi and Taddeo, 2016).

Understanding citizen perspectives is pivotal for shaping inclusive, ethically sound, and socially acceptable AI deployment strategies for AI in security and policing. Citizens' concerns, and preferences play a fundamental role in determining legitimacy, trustworthiness, and societal acceptance of AI systems used by law enforcement and security agencies (Crawford and Calo, 2016). Thus, this article aims to fill the gap in existing discourse by elucidating the significance of citizen perspectives in defining the preferred ownership of AI tools

within security domains. By amplifying these insights, this study underscores the importance of inclusive governance frameworks that prioritize the amalgamation of citizen opinions, ensuring responsible and beneficial AI integration into society.

1.1 AI Ownership in Security and Policing Contexts

The deployment of AI tools within the security domain such as predictive policing algorithms (e.g., PredPol) and facial recognition systems (e.g., Clearview AI) highlight the complexities of AI ownership and showcase the intersection of technological innovation, legal frameworks, and societal implications. The discourse on AI ownership revolves around control, accountability, and responsibility for actions and decisions executed by these intelligent systems. Floridi (2019) argues that ownership extends beyond possession to include responsibility for AI actions, such as biases, errors, and ethical implications. This aligns with broader discussions on accountability and the necessity for transparency and oversight in AI deployment (Kroll et al., 2017).

The adoption of AI technologies such as automated license plate readers (ALPRs) and crime mapping tools, raises both concerns and opportunities. The potential of AI to augment law enforcement capabilities and optimize resource allocation is juxtaposed with apprehensions regarding privacy infringement, fears of biases, and the erosion of discretion in decision-making (Aloisi and Gramano, 2020; Mittelstadt et al, 2016). The ownership and deployment of these tools by police raises debates around balancing security with societal values and whether ownership should be exclusive to law enforcement or more distributed (Orwell, 2000).

1.2 Aim of This Study

This research aims to explore the perspectives of citizens regarding the preferred owner(s) of AI tools for policing and law enforcement applications. Through semi-structured interviews, this study seeks to elucidate public perceptions, concerns, and preferences concerning AI ownership, contributing to informing policy frameworks and ethical guidelines governing the deployment and ownership of AI tools in security domains. Understanding citizens' perspectives towards ownership of AI tools is crucial, especially given the legal and moral implications involved (Robaey, 2015; Hayes et al, 2020). By understanding the rationale behind citizens' viewpoints on access and ownership of AI policing capabilities, this study aims to contribute to the theoretical and social context in which security opportunities align with community needs and perspectives, leading to potential endorsement of a virtuous implementation of AI within policing.

2. Methodology

Semi-structured interviews were conducted in eight different countries (UK, Netherlands, Italy, Spain, Portugal, Czech Republic, Germany, and Greece), involving 111 participants. Participants were recruited based on specific group specifications relevant to each of the partner countries as part of the AIDA2020[1] joint project. The interviews focused on citizens' attitudes towards AI use by law enforcement agencies (LEAs), and namely ownership. Interviews were chosen as a qualitative approach to better understand and integrate citizens' perceptions towards AI ownership, following a structured theme of scenario-based interview questions. The data collected from these interviews underwent thematic and content analysis to identify main themes and patterns. Participant responses were coded and clustered into high-order categories, allowing for the emergence of common perspectives reflecting preferences for AI ownership in different contexts. This analysis, performed using NVivo's qualitative data analysis software, enabled the exploration of citizen perspectives on AI ownership, contributing to the broader discourse on AI ownership within security contexts.

2.1 Participants

Participants were recruited by researchers in the eight participating countries. The selection allowed free choice of the citizen group to allow partner countries to choose groups that they considered relevant and of interest in their national context. A total of 111 individuals participated. Germany focused on young women (18-25 years, n=16), Czech Republic (n=10) focused on young people in general between 18-25 years old, while Italy addressed older citizens (65+ years). The Netherlands focused on expatriates (n=16) with experience in the Cybersecurity field. In the UK, 11 individuals with a migration background were recruited. Spain (n=20) and Portugal (n=6) chose participants who are familiar with AI, while Greece (n=16) decided to adopt an open

[1]AIDA2020: Artificial Intelligence and Advanced Data Analytics for Law Enforcement Agencies. https://www.project-aida.eu/

selection of participants of diverse occupations and disciplines. Table 1 shows the demographic and gender distributions of the selected groups.

Table 1: Demographic characteristics of participants per country

Country	Number of Participants	Group specifications	Gender distribution Women / Men	Average age
Czech Republic (CZ)	10	Young people between 18-25 years old	60% / 40%	23.4
Germany (DE)	16	Young women between 18-25 years old	100% / 0%	26.3
Greece (GR)	16	Experts in IT Law and IP law	78.57% / 24.4%	33.25
Italy* (IT)	16	Older citizens (65+)	50% / 50%	72.8
Netherlands (NL)	16	Expatriates	43.7% / 56.2%	26.3
Portugal (PT)	6	People with limited knowledge of AI	50% / 50%	44.7
Spain (SP)	20	AI experts	40% / 60%	37.5
UK	11	Citizens with migration background	72.7% / 27.2%	33.4
Total	*111*		*62.2% / 37.8%*	*38.3*

2.2 Data Collection

A total of 111 semi-structured interviews were conducted addressing citizens' attitudes towards AI use by LEAs. While the first part of the interview focused on overall questions of acceptance and acceptance conditions, the second part offered participants the opportunity to reflect more specifically on potential ethical dilemmas of AI use and possible resistance. This paper focuses on one part of the interview, particularly the part where participants were asked about whether AI capabilities should be limited to police only, or would they want to have access to such tools themselves.

The interviews were conducted either face-to-face or online by researchers in each participating country to allow participants to react to questions in their own language. Therefore, all participating countries provided participants with the information sheet, the informed consent, and the interview guidelines in the language of the respective country, except for the Netherlands where the researchers chose to share the documents with the participants in English instead of Dutch since they interviewed expatriates. Follow-up questions, prompts and comments were made by interviewers in each country to encourage participants to elaborate on the rationales for their choices. All interviews were audio-recorded. Some were transcribed as summaries, others were transcribed verbatim, and all the data was anonymized before analysis.

2.3 Data Analysis

The transcripts and the summaries obtained in the native countries' languages were translated to English using a designated translation software which was followed by close proof-reading. The English transcripts/ summaries were used for the data analysis. Our analytic approach followed thematic (Auerbach and Silverstein, 2003) and content analysis principles (Krippendorff, 2004) for the purpose of identifying main themes and patterns in the data. First, the answers were coded in cycles, starting with open or initial coding (Charmaz, 2006), followed by clustering into high-order categories for each main topic. Thematic analysis was used to allow for the thorough evaluation of the statements made by each participant which revealed common perspectives whereby specific ownership was preferred by participants, depending on the situation, and was justified by different rationales. This coding was performed using NVivo's qualitative data analysis computer software package. As a second step, participant's responses were thoroughly reviewed, coded, and assessed for similarities, then clustered under common sub-themes. This process is largely exploratory,

whereby the analysis does not rely on any predefined categories or features in creating the clustered perspectives.

2.4 Ethics

This study has been approved by the ethics committee of the authors' affiliated university. Additionally, participants were informed of the context and legal basis of the study, the details of data handling and their rights through the information sheet and informed consent form which the participants had to sign prior to the interview. The right to withdraw and to opt out from providing demographic information was also explicitly stated in the above-mentioned forms. All data was analysed in pseudonymized form.

3. Results

3.1 Analysis of Perspectives

The approach revealed five disparate perspectives towards the preferred ownership of AI capabilities for security. Figure 1 shows the percentages of respondents who favoured each form of AI owner. Below, we provide an in-depth analysis of these responses, citing participant comments in italics to clarify their decision-making process. Each perspective is summarized with a descriptive title highlighting key aspects.

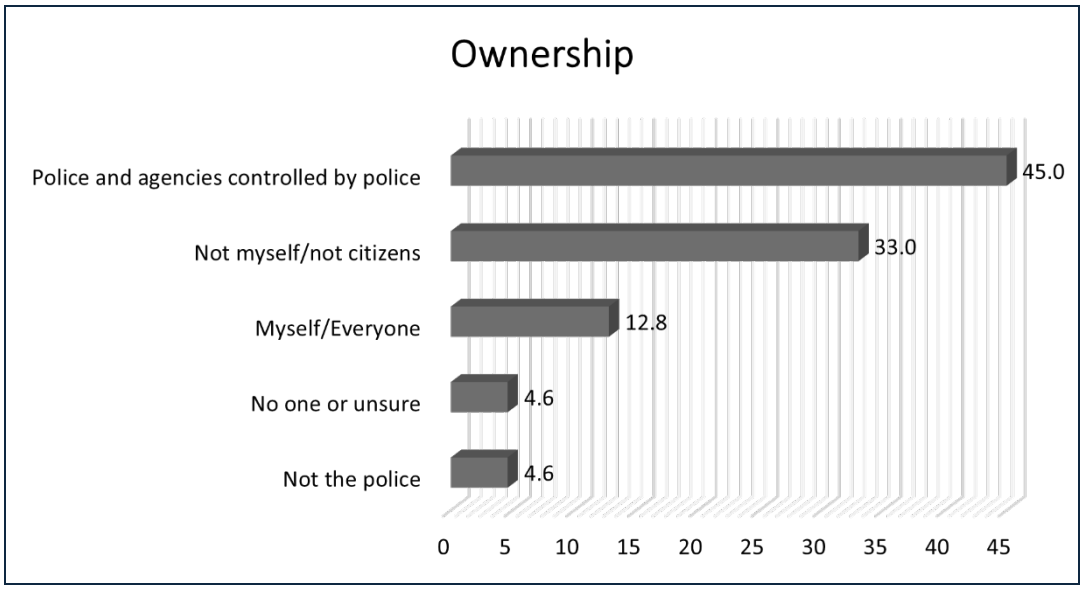


Figure 1: Clustered Perspectives on preferred owner(s) of security-related AI (in percentages of participants in the sample)

Perspective 1: Preference for Police, LEAs, and Government Agencies

This was the most common perspective, with most participants preferring LEAs/Police to own and control AI tools. They cited public safeguarding, reducing costs, and proactively preventing crimes as key reasons. The trust in LEAs stems from their established role for safeguarding society and their existing access to confidential information, coupled with their training to handle extreme cases and the biases of AI.

Participants believe that police and LEAs have the necessary skills, competence, and ethical obligations to use AI responsibly compared to private entities. Participants specifically emphasized the role of police in protecting citizens: "This must be with the police; they must protect us" (PT-05). Concerns about private entities mishandling data surpassed worries about police use of personal data, reflecting a general acceptance of the police's role in safeguarding. They trust police forces to protect private information and use AI for public safety. As NL-13 said, "I trust the police with this kind of information."

Some participants suggested extending AI ownership to other government agencies, intelligence units, independent auditors, academia, and corporations with appropriate vetting and accountability measures. Centralized authorization under LEA supervision was also suggested, reflecting the belief that protecting the population is a core police mission (PT-05).

Perspective 2: Preference against Police Ownership

A smaller number of participants opposed police ownership of AI, citing the need for broader monitoring and evaluation to ensure effective use. They doubted the police's expertise and transparency in handling AI for crime prevention. GR-12 remarked, “For police I don’t know how they will use it, and I am a bit overwhelmed knowing that.” This group thus represents an opposing view to Perspective 1. The two opposing perspectives illustrate the importance of trust in police and perceptions of competence as basis for AI ownership.

Perspective 3: Preference for no ownership by citizens (including themselves)

This perspective indicates a general rejection of AI ownership by citizens, citing their lack of qualifications to handle such tools. This included participants themselves as well as others. As IT-11 stated, “It would be a waste of time if I had some computer tools to protect myself because I wouldn't be able to handle them, honestly.”

Others felt unsafe with such information, preferring police to handle AI for protection (PT-04). Participants compared AI ownership to gun ownership, fearing misuse: “I don't think it should be something that goes out to just anyone because that turns out to be a bit like gun ownership” (PT-01). They were concerned about AI becoming a weapon in the wrong hands (NL-04).

Perspective 4: Preference for Everyone (including themselves)

Contrary to Perspective 3, some participants supported citizen ownership of AI tools for personal safeguarding, especially against cybercrimes. As IT-4 stated, “For cybercrime, I think that all of the people that wanted something to protect themselves from cybercrime should be allowed to have those.” However, some suggested limiting AI ownership to themselves to avoid misuse by others (GR-16). This perspective thus suggests that citizen ownership may be warranted for very specific purposes, while also expressing fear of mistrust in other citizens’ intentions.

Perspective 5: Preference for No one to own AI / unsure about preferred Owner

The final perspective argued against anyone owning AI tools due to concerns about data accuracy, algorithm transparency, and potential misuse (SP-03). Some participants expressed uncertainty about potential AI owners and preferred not to elaborate. This perspective indicates generalised concerns about the viability of AI for security, which translated into a position that no one should own such tools.

3.2 Interpretation of Viewpoints

Participants from various countries shared similar reactions toward police use of AI, though there were notable differences in reasoning. Supporters of perspective 1 believed that only police should own AI tools, citing their role in public safety and trustworthiness with personal data. Some also supported AI use by other LEAs and government bodies for better oversight and objective evaluation. Conversely, participants in perspective 2 opposed police use of AI due to concerns around their capability to manage AI technologies and the transparency of AI-driven decisions. Perspective 3 revealed strong opposition to AI ownership by the public in fear of inadequate knowledge and training, linking it to the risks of public gun ownership in the U.S. In contrast, perspective 4 advocated for public access to AI tools, provided proper training and ethical guidelines are in place. These participants emphasized the need for public involvement in evaluating AI use in policing to ensure transparency. Perspective 5 included a minor group entirely opposed to AI use by anyone, preferring traditional non-AI technologies that have been effective so far.

4. Discussion

In our contemporary world, ownership spans both material items and intellectual properties, with laws protecting these rights to ensure owners can control and benefit from them (Hayes et al., 2020). Ownership of AI includes the rights to possess, use, manage, and benefit from these tools, along with associated responsibilities (Robaey, 2015; Honoré, 1961). The prevalence of AI tools necessitates exploring their ownership to ensure adherence to human rights and privacy laws, which could enhance societal acceptance of AI in policing (Ezzeddine et al., 2022).

This paper examines citizens' views on AI tools’ ownership, discussing these perspectives in relation to debates around AI governance in policing. The findings touch on legal and ethical implications, including roles, responsibilities, expertise, accountability, and public acceptance (Carrasco et al., 2019; Neudert et al., 2020), as well as public willingness to trade privacy for safety (Pavone & Esposti, 2012).

Participants' preferred AI owner correlated with roles, responsibilities, and benefits to the public. Many expressed frustrations over data used for personalized ads without consent, while fewer were concerned about police accessing the same data. Significant concerns included AI's lack of autonomous moral operation and biases in police decision-making (Farina et al., 2020). This correlates to ongoing ethical discussions and recent research highlighting the importance of ownership models that prioritize ethical principles and societal values to foster acceptance (Nemitz, 2018).

Opinions on police ownership were divided. Some trusted police ownership due to their safeguarding roles, while others doubted police expertise and transparency. Some even suggested extending AI ownership to other government and professional entities for oversight (Martin, 2019). This debate aligns with broader discussions around transparency and need for police digitalization strategies (Gundhus et al., 2022). It also reiterates the importance of trust in LEAs and its significant impact for public support of AI in policing. In this context, participants specifically highlighted the importance of police legitimacy and accountability in accessing personal data, reflecting expectations that police actions should prioritize national security (Tyler & Huo, 2002).

Moreover, participants' views often balanced privacy and security. Some accepted police AI use but mistrusted others, suggesting stringent regulations to prevent misuse (Jones & Haggerty, 2021). The debate on privacy versus safety remains critical, with calls for ethical considerations in AI deployment (Lyon, 2002; DiVaio et al., 2022). Some saw AI in surveillance as potentially invasive, linking it to mass data collection and bulk data analysis (Albrecht, 2020). This contrasts with historical arguments defending surveillance for improved security (Bentham, 1791) by highlighting that excessive monitoring could reduce trust in law enforcement (Yesberg et al., 2021).

Participants also recognized the significant influence of corporations on AI development and regulation. They were more critical of AI in targeted advertising by corporations than to its use by police, citing trust in law enforcement's accountability and safeguarding principles (Ezzeddine et al., 2022).

In terms of practical implications, the findings suggest options for differentiated ownership models, as well as the need for robust legal and ethics oversight, and community engagement to address preferences for AI tool ownership in law enforcement. These measures could enhance transparency, trust, and shared responsibility, aligning AI use with societal values and priorities. Below we list some concrete options to increase citizen support of AI use for security purposes:

- Differentiated Ownership Models: Exploring disparate ownership models involving law enforcement and citizen representatives (e.g., acknowledging data origins and dependencies in safety production) to enhance transparency and trust. This could include shared ownership models and public-private partnerships to ensure no single entity dominates (Crawford et al., 2019; Eubanks, 2018). Ownership should reflect data origins and AI's impact on communities, ensuring those most affected have a say in decision-making (O'Neil, 2016).
- Robust legal and Ethical Oversight: Implementing stringent ethics guidelines and independent oversight mechanisms to ensure compliance and prevent misuse. This involves developing comprehensive legal guidance and ethics standards focusing on privacy, bias mitigation, and accountability, with input from diverse stakeholders (Floridi, 2019). Alternatively, independent bodies auditing AI tools can ensure compliance with legal, professional and ethics standards through regular reports (Whittaker et al., 2018).
- Community Engagement: Promoting community-led forums for citizen input in AI tool ownership and control decisions. This includes facilitating citizen assemblies and public consultations to ensure community input is integrated into AI governance, aligning it with public values (Zuboff, 2019). Additionally, creating community-led oversight committees to monitor AI use and advocate for necessary changes can promote education and transparency (Benjamin, 2019).

In summary, the study highlights the need for differentiated considerations on AI ownership and deployment and the importance of citizen engagement to ensure trust and accountability. Future research should explore deeper rationalizations around AI ownership preferences, focusing on roles, responsibilities, expertise, accountability, and the balance between costs and benefits.

As for limitations, our sample was skewed younger and included more women than men, suggesting the need for broader participant demographics in future research. Additionally, some participants felt unqualified to comment on AI ownership, indicating a need for inclusive discussions accounting for all perspectives. Future

studies should address these issues to provide a comprehensive understanding of public perceptions and factors influencing preferences for AI ownership in policing.

5. Conclusion

This study critically examines citizens' perspectives on legitimate ownership of AI technologies, emphasizing the balance between rights and duties of AI implementers, particularly in policing and security. It suggests a need for public involvement in AI tool implementation, accountability, and transparency in data processing and decision-making (Vestby and Vestby, 2019). Moral responsibilities are highlighted, with citizens seen as potential owners, stressing the need for ethical governance and trust in AI applications (Lawrence et al., 2018). The research suggests active public involvement in decision-making to align ownership structures with societal values and ethical considerations (Pavone and Esposti, 2012). This approach enhances understanding of AI acceptance and trust, emphasizing inclusive governance and ethical frameworks (Benjamin, 2020; Ferguson, 2017).

Our findings align with ethical AI principles proposed by entities like the European Commission, addressing public concerns about accountability, legitimacy, and privacy (Ezzeddine et al., 2022). They underscore the importance of training, skills, and ethical principles for AI regulation, regardless of ownership (Albrecht, 2020). Citizens in the study exhibited diverse, instance-based perspectives on AI in policing, guided by roles, responsibilities, and a balance of costs versus benefits, rather than outright rejection (Angwin et al., 2016). This diversity indicates the need for differentiated communication and engagement with citizens about the deployment of AI capabilities for security that acknowledges multiple owners – not only of AI but also of the responsibility to secure society (Bayerl et al., 2022; Terpstra, 2009). By involving the public and ensuring transparency, LEAs can integrate AI technologies while maintaining ethical standards and public trust, reflecting citizens' inquisitive mindset towards ethically guided AI deployments (Yesberg et al., 2021).

Acknowledgement

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 883569 as part of the AIDA project (AIDA - Artificial Intelligence and advanced Data Analytics for Law Enforcement Agencies). For the purpose of open access, the authors applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

Albrecht, H.J. (2020). Data, Data Banks and Security. European Journal of Security Research, 5(1), pp. 5-23. doi:10.1007/s41125-019-00062-9.

Aloisi, A. and Gramano, E. (2020). Artificial Intelligence is Watching You at Work. Digital Surveillance, Employee Monitoring and Regulatory Issues in the EU Context. Comparative Labor Law and Policy Journal, pp. 95-121.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks. [online] Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed 15 February 2022].

Auerbach, C.F. and Silverstein, L.B. (2003). Qualitative Data: An Introduction to Coding and Analysis. NYU Press.

Bayerl, P.S., Butot, V., & Jacobs, G. (2022). Produktion urbaner Sicherheit aus Bürgerperspektive. In: D. Wehe, H. Siller (eds.), Handbuch Polizeimanagement, Springer Nature, pp. 1-18.

Benjamin, G. (2020). Facial recognition is spreading faster than you realize. The Conversation. [online] Available at: https://theconversation.com/facial-recognition-is-spreading-faster-than-you-realise-132047 [Accessed 15 February 2022].

Benjamin, R. (2019) Race After Technology: Abolitionist Tools for the New Jim Code, Polity.

Bentham, J. (1791). Panopticon, or, The Inspection-House. Dublin: T. Payne.

Carrasco, M., Mills, S., Whybrew, A. and Jura, A. (2019). The Citizen's Perspective on the Use of AI in Government. BCG Digital Government Benchmark. [online] Available at: https://www.bcg.com/publications/2019/citizen-perspective-use-artificial-intelligence-government-digital-benchmarking.aspx [Accessed 15 February 2022].

Charmaz, K. (2006). Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. London: Sage Publications.

Chohan, S.R. and Hu, G. (2020). Strengthening digital inclusion in e-government: Cohesive ICT training programs intensify digital competency. Information Technology for Development, 28(1), pp. 1-23. doi:10.1080/02681102.2020.1841713.

Crawford, K. and Calo, R. (2016). There is a blind spot in AI research. Nature, 538(7625), pp. 311-313.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E. and Whittaker, M. (2019) AI Now 2019 Report, AI Now Institute.

Di Vaio, A., Hassan, R. and Alavoine, C. (2022). Data intelligence analytics: A bibliometric analysis of human–AI interaction in public sector decision-making effectiveness. Technological Forecasting and Social Change, 174, 121201. doi:10.1016/j.techfore.2021.121201.

Eubanks, V. (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, St. Martin's Press.

European Commission. (2020). EU Economic and Social Committee Regulation on AI in Europe, 237 final. Brussels: European Commission. [online] Available at: https://www.njb.nl/umbraco/uploads/2018/7/COM-2018-237-F1-EN-MAIN-PART-1.PDF [Accessed 18 February 2022].

Ezzeddine, Y., Bayerl, P.S. and Gibson, H. (2022). Citizen Perspectives on Necessary Safeguards for the Use of AI by Law Enforcement Agencies. arXiv.org. doi:10.48550/arXiv.2306.01786.

Farina, M., Zhdanov, P., Karimov, A. and Lavazza, A. (2022). AI and society: A virtue ethics approach. AI & Society. doi:10.1007/s00146-022-01545-5.

Ferguson, A.G. (2017). Policing Predictive Policing. Washington University Law Review, 94(5), pp. 1109-1189.

Floridi, L. (2019). The Logic of Information: A Theory of Philosophy as Conceptual Design. Oxford University Press.

Floridi, L. and Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A, 374(2083), 20160360.

Gundhus, H.O., Talberg, N. and Wathne, C.T. (2022). From discretion to standardization: Digitalization in police organizations. International Journal of Police Science & Management, 24(1), pp. 27-41. doi:10.1177/14613557211036554.

Hayes, P., van de Poel, I. and Steen, M. (2020). Algorithms, values, and justice in security. AI & Society, 35, pp. 533-555. doi:10.1007/s00146-019-00932-9.

Jones, R. and Haggerty, K.D. (2021). AI policing: A white paper. Surveillance & Society, 19(3), pp. 331-337.

Krippendorff, K. (2004). Content Analysis: An Introduction to its Methodology. Thousand Oaks; London; New Delhi: Sage Publications.

Kroll, J.A., et al. (2017). Accountable Algorithms. University of Pennsylvania Law Review, 165(3), pp. 633-705.

Lawrence, D., Peterson, B. and Thompson, P. (2018). Community views on Milwaukee's police body-worn camera program. Justice Policy Center, Urban Institute.

Lyon, D. (2002). Everyday surveillance: Personal data and social classifications. Information, Communication & Society, 5(2), pp. 242-257. doi:10.1080/13691180210130806.

Martin, G. (2019). Public attitudes towards police use of facial recognition technology. Police Quarterly, 22(3), pp. 349-368.

Mittelstadt, B.D., et al. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180089. doi:10.1098/rsta.2018.0089.

Neudert, L.M., Knuutila, A. and Howard, P. (2020). Global Attitudes Towards AI, Machine Learning & Automated Decision Making: Implications for Public Service and Good Governance. University of Oxford: Oxford Commission on AI and Good Governance.

O'Neil, C. (2016) Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown.

Orwell, G. (2000). 1984 Nineteen Eighty-Four. Introduced by Thomas Pynchon. England: Penguin Classics.

Pavone, V. and Esposti, S. (2012). Public assessment of new surveillance-oriented security technologies: Beyond the trade-off between privacy and security. Public Understanding of Science, 21(5), pp. 556-572. doi:10.1177/0963662510376886.

Peeters, R. and Widlak, A. (2018). The digital cage: Administrative exclusion through information architecture – The case of the Dutch civil registry's master data management. Government Information Quarterly, 35(2), pp. 175-183.

Robaey, Z. (2015). Looking at moral responsibility for ownership: A way to deal with hazards of GMOs. Journal of Agricultural and Environmental Ethics, 28(1), pp. 43-56. doi:10.1007/s10806-014-9517-8.

Schuilenburg, M. and Peeters, R. (2020). Algorithmic society. Algorithmic Society, pp. 1-15. doi:10.4324/9780429261404-1.

Terpstra, J. (2009a). Citizen involvement in local security networks. Security Journal, 22, 156–169.

Tyler, T.R. and Huo, Y.J. (2002). Trust in the Law: Encouraging Public Cooperation with the Police and Courts. Russell Sage Foundation.

Vestby, A. and Vestby, J. (2019). Machine Learning and the Police: Asking the Right Questions. Policing: A Journal of Policy and Practice. doi:10.1093/police/paz035.

Whittaker, M., Crawford, K., Dobbe, R. and Fried, G. (2018) AI Now 2018 Report, AI Now Institute.

Yesberg, J., Brunton-Smith, I. and Bradford, B. (2021). Police visibility, trust in police fairness, and collective efficacy: A multilevel structural equation model. European Journal of Criminology. doi:10.1177/14773708211035306.

Zuboff, S. (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, PublicAffairs.

How Is Fake News Spread? An Analysis of the Dissemination Process: Actors, Channels, and Motives

Anastasiia Iufereva
IULM University, Milan, Italy

anastasiia.iufereva@studenti.iulm.it

Abstract: Even though fake news is widely recognized as one of the most serious threats in the post-truth era, there are still some gaps regarding fake news dissemination. To address these gaps, it is necessary to conduct theoretical research revising and analyzing the latest scientific developments on this topic paying attention to social, psychological, and technological contexts in which fake news is constructed and spread. Understanding this process might help to improve the media literacy of users by raising awareness and forming critical attitudes to by-products of the information environment. This article presents a literature review on fake news dissemination based on an analysis of 106 papers extracted from the Web of Science, Scopus, and Google Scholar databases.The author focuses on identifying the main actors who spread fake news. Additionally, a typology is proposed to organize knowledge about the motives behind creating fake news, based on social, psychological, and cognitive factors.

Keywords: Fake news, Deed fake news, Disinformation, Dissemination, Social media

1. Introduction

The issue of fake news is by far one of the most acute problems of the post-truth era, not only at a practical level when the fact-checking process requires particular attention to the facts but also at a conceptual level. The latter, in turn, is determined by the difficulty that arises from the definition of the term fake news, as no generally accepted definition has yet been proposed. Generally, fake news can be considered as a publication circulated online (see Rocha et al., 2021; Bakir & McStay, 2017) that is completely or partially fictitious, dealing with real events and situations. To establish credibility, fake news adopts such a journalistic format as news, employing titles, images, and content designed to capture readers' attention, inspire confidence, and consequently mislead them. Fake news is always deliberately generated to manipulate the targeted audience's opinion for some reason. To achieve this goal different platforms, algorithms, and approaches are utilized. Therefore, this paper will focus on the sources of fake news and the motives of people for spreading fake news.

2. Methodology

For this research, literature analysis as a qualitative methodology has been used to interpret the existing body of research on the topic. First, the author identified sources: 106 papers related to fake news were collected from the Web of Science, Scopus, and Google Scholar databases. Most of the articles were published after 2016, a period when "fake news" gained significant prominence and was even named Word of the Year by Collins (since 2016, the usage of the term has surged by 365%). Then, a comprehensive reading of the collected sources was organized, focusing on two main areas: (1) the participants involved in spreading fake news, and (2) the primary motives behind its dissemination. The author considered the arguments, methodologies, and findings on these aspects and combined insights from different sources to identify key trends and gaps in the research. Finally, the author used techniques such as interpretation and categorization to expand existing knowledge and address gaps in the dissemination process of fake news, forming a comprehensive view of the actors, channels, and motives in a coherent and organized manner.

3. Literature Analysis Results

3.1 Actors: Real Human Beings or Non-Humans

Zhang and Ghorbani (2018) suggested two main actors: real humans and non-real humans.

Real humans are crucial in fake news dissemination. These individuals can be categorized into two groups: malicious and benign. The former deliberately share content recognized as false for different motives: political or ideological (Di Domenico et al., 2021). Some malicious users are paid to disseminate fake news (Zannettou et al., 2019). The intentional generation of fake content by malicious online users makes it challenging to distinguish between fake and true information based solely on content and linguistic analysis (Shu et al., 2017). Moreover, benign online users contribute to the spread of fake news by sharing fake news across social media platforms, believing it to be true.

Non-real humans indirectly contribute to the spread of fake news as carriers on the Internet. These automated accounts are programmed by real humans to disseminate fake news. Whether fake news is spread manually or automatically, real humans aiming to disrupt the credibility of online social communities are the ultimate creators (Zhang& Ghorbani, 2018).

3.2 Spreading Channels Features

There are different types of accounts and platforms used for spreading fake news. This research consolidates all currently used channels and platforms for fake news dissemination, categorizing them as follows:

Social Bots: common non-human agents exist in the form of social bots, involving computer algorithms that are designed to exhibit human-like behavior, artificially produce content, and interact with humans on social media (Ferrara et al., 2016). By artificially performing actions such as liking, sharing, and commenting, social bots are known to accelerate the spread of fake news (Lazer et al., 2018, Shao et al., 2018). Social bots play a significant role in spreading fake news on social media for two reasons. First, they amplify fake news in the early spreading moments and spearhead the fake narrative to go viral (Azzimonti & Fernandes, 2018). Second, they can recognize and target influential users through replies and mentions, so influential users are involved in the spreading process (Shao et al., 2017). Social bots can perform a wide range of activities and have become increasingly sophisticated or 'smart' since they can scout the Internet for information to fill their profiles and post collected materials at a scheduled rate (Ferrara et al., 2016). Given the vulnerability of social media to large-scale infiltrations, social bots are often organized in networks, resulting in hundreds of unique accounts being managed by only one computer (Boshmaf et al., 2011).

Cyborgs. If the social bot refers to computer software that generates messages automatically, the cyborg refers either to a bot-assisted human or a human-assisted bot. It means that these media accounts are not always bots and are not always people. As a rule, cyborg accounts, being registered by humans as camouflage, start to disseminate information (e.g., post tweets) and communicate with the social community, blending automated activities with human input. "They are a hybrid of both: accounts that are detected as bots in one time window but identified as humans in another" (Shu et al., 2017).

Trolls are real human agents. They work to disrupt online communities, provoking people into an emotional response, and hindering mature discussions (Kim et al., 2021). Trolling behaviors are highly affected by people's moods and the context of online discussions, which enables the easy dissemination of fake news among otherwise "normal" online communities (Shu et al., 2017). For instance, trolls may provoke discussion on a particular topic in the comments to the news (most often, fake news), spreading false information and inciting conflicts between users. Active discussion brings this news to the "top," resulting in the news appearing in the feeds of many users. According to social media algorithms, high activity in comments (positive or negative) increases the relevance and popularity of the news overall. Furthermore, troll armies might be active on social media to advance political agendas and cause social turmoil. They also might spread fake news, significantly impacting people's beliefs.

Fake news media is a website that contains false information intended to mislead readers. The main goal of such resources is to increase traffic and, as a result, gain financial profit or political benefits. These news media exist in two forms. First, some web pages often resemble legitimate news organizations because of their structure, but they are established purposely to publish fake news (Zhang& Ghorbani, 2018). Second, some websites are portrayed to be spinoffs of other legitimate news sites with high reputations. They are structured in such a way (style, visual design, etc.) to make visitors believe they are visiting major news outlets like BBC News or The Sun (Gilbert, 2016). Tandoc and Albright (2016) note that fake news sites regularly rely on an "ecosystem of real-time propaganda." Both forms of these websites might be managed by real users or networks of bots that automatically disseminate fake news on these websites and social media (Tandoc & Albright, 2016). "This gives the reader the sense that many others are also reading and liking the item, and if they go to another site to verify the item, they will likely find it there, again adding a veneer of legitimacy to the piece" (Edson et al., p. 144, 2017).

Search engines and aggregators: search engines, such as Google Search and Yahoo, have become crucial intermediaries to guide users to online news articles. For example, the Google search engine accounts for an average of 30% of traffic to news websites (Olmstead et al., 2017). In recent years, more than 99% of respondents surveyed in 2017 used Google Search to find information at least once a day (Finkel et al., 2017). Search engines that provide news are multi-sided markets because they create a space for the interaction between media producers and readers. Search engines generate an algorithm-driven search ranking of articles

from which consumers can choose. However, ranking articles lead to a situation where frequently read articles get to the top and therefore become more popular compared to lower-ranked articles because they are rarely read. Among such articles might be fake news, as proven in the report Fake News and Misinformation (Finkel et al., 2017). It was noted that Google Search plays a critical role in the misinformation landscape spreading fake news. Companies pay Google to place the companies' advertisements on relevant search results. This system allows malicious websites that spread fake news get the possibility to be seen by users.

Deepfake news: In this section, attention is paid to using artificial intelligence to create fake news. These non-human agents are managed by humans who utilize a wide range of tools suggested by various websites on the Internet to develop media products called "deepfake news" which is considered as the next generation of fake news (see Nguyen et al., 2022). The term combines "deep learning" and "fake" – deepfakes rely on artificial intelligence and machine learning. This phenomenon refers to the use of artificial intelligence to alter images and photos in multiple directions to manipulate public opinion. The appearance of this type of material was due to social and technological factors. Last few years it is noticed the popularity of images and videos that are shared by users on social media frequently. Hence, there is a tendency to widespread different formats of presentation of this visual content as reels on Instagram or shorts on YouTube. The high popularity of such types of consumption content led to the creation of malicious tools that are designed to manipulate these materials altering them for mercenary purposes.

3.3 Motives for Creating and Spreading Fake News

3.3.1 Motives for creating fake news: theoretical grounds

First, to understand the nature of the process of creating fake news by users, it is necessary to consider the social and economic science foundations that describe the intrinsic motives of people creating news and influencing the media ecosystem. To describe the social dynamics that contribute to the proliferation of fake news, this paper refers to the rational theory of fake news interactions. "It is modeled from an economic game theoretical perspective by formulating the news generation and consumption cycle as a two-player strategy game" (Shu et al., 2017).

To explain fake news, this research assumes that the publisher plays a key role in the information ecosystem, and this media environment can benefit them in two ways: (1) short-term utility: the incentive to maximize profit, which is positively correlated with the number of consumers reached; and (2) long-term utility: their reputation in terms of news authenticity. Fake news occurs when the short-term utility dominates, and there are several aspects to this: (i) publishers who create fake news think in terms of the short-term perspective, which also explains why the life expectancy of fake news is short; (ii) there is a desire to maximize profit, which drives the publisher to make the fake news more attractive to the public (e.g., the author focuses on "hot" topics in politics, economics, and social life, since fake news is mainly generated around the most discussed topics, which are relevant for search queries). Various tools might be used: flashy headlines, bias, and changes to the intended or true meaning of something, etc.

At the same time, it is difficult to predict when fake news will become more popular, because a high level of distortion does not guarantee the rise in popularity or profit. However, there is a direct correlation between profit margin and the number of people who see and share the fake news.

3.3.2 Types of motives for creating fake news

Most studies identify a few crucial motives for the creation and dissemination of fake news as political and financial. According to our perspectives, motives are a great deal more profound and might relate to other social dimensions like entertainment, fear and panic, social status, and prestige.

It has been suggested that the fake news genre is created deliberately to deceive. This intention can be seen as a defining element of fake news. For this reason, the author will not consider misinformation based on religious, social, or political bias when a person genuinely believes it. Unintentionally presenting false information based on bias and personal beliefs is not considered fake news but rather misinformation (e.g., errors that journalists unintentionally make).

Political intention

This intention has been examined extensively in highly cited research (see Allcott & Gentzkow, 2017; Bakir & McStay, 2018; Lazer et al., 2018; Cheuk Hang Au, 2018; Weir, 2009). Based on the analysis of these papers, it is

necessary to highlight the following points. First, fake news is designed to get people to change their minds about their political beliefs and uses falsehoods to manipulate public opinion quickly and across a wide audience. Rather than delivering arguments to persuade people to adopt new beliefs, it uses false events to turn people away from existing beliefs (Townsend, 2017). Fake news might be created to support or promote political figures, political movements, social policies, political ideologies, or election candidates, often involving discrediting or undermining those on the opposing side. Historically and currently, political actors may disseminate false information to gain public support. This negatively affects people by reinforcing and worsening in-group biases. Hence, a major impact of fake news is ideological polarization, a condition under which opinions on an issue are opposed to some theoretical maximum, and polarization is a process that refers to the increase in such opposition over time (Cheuk Hang Au, 2020). In highly ideologically polarized contexts, users tend to confirm their pre-existing views by communicating with others with similar views inside a closed system, rather than attempting to understand the complete picture by appealing to alternative opinions or convictions. As a result, people may become more radical, avoiding exposure to opposing viewpoints and preferring to stay inside a closed system with limited alternative opinions or convictions. Ideological polarization may reinforce existing cognitive biases, while the digital media environment may only foster this effect. Furthermore, news feeds on social media algorithms usually create an environment that is comfortable for us by suggesting and showing us like-minded information. According to Cheuk Hang Au (2020), when political issues are not too divisive, people tend to be more accuracy-oriented when looking for relevant information. However, "when the issue has received substantial attention, people may be more likely to engage in biased information search and processing, and hidden conflicts in values between each other may become more visible" (Cheuk Hang Au, p.12, 2020). As a result of ideological polarization, the diversity of opinions and arguments may be compromised. People in extremely homogenous groups may ignore facts that would prove their arguments wrong, and the dissemination of online disinformation may increase.

Financial benefit

Financial benefit is another strong motive for fake news creators (Rochlin, 2017; Kasra, 2017). Nowadays, on the Internet, two formats of disinformation compete for the right to make a profit: clickbait and fake news. Despite having a common financial goal, it is necessary to pay attention to the differences between them. First, clickbait is a type of content that includes sensational headlines that encourage visitors to click on a link to a particular web page that usually does not have satisfactory content (usually it looks like advertisement-like content or just a headline and short description of a fabricated event published on suspicious websites that do not pretend to be reliable mass media sites or masquerade as legitimate news publishers). On the contrary, fake news mimics the news format and is not usually perceived as second-rate material. It is comprised of a headline and body of text that are both misleading and composed of fabricated facts. Fake news containing a headline and body text as the main elements of this format can be posted on social media. Fake news may contain links to websites that either masquerade as legitimate but are fake, or to sites with advertising content. For example, media publishers that mimic a prominent one may create fake news to gain more readers and attract more clicks across social media (Allcott & Gentzkow, 2017). Spammers also can make money by publishing fake news on social media that has the form of legitimate news publishers and leads to websites, which are often mostly ads. In both cases, social media manipulation can be used to drive traffic to a particular site. Hence, creators of fake news can draw significant advertising revenue when users click and go to the website since many conversions to a site are the crucial reason for advertisers to place advertisements on this site. Another example is how fake news regulates stock prices in financial markets. Kogan et al. (2021) show that fake news is a "part of a pump-and-dump scheme to defraud securities markets. (…) fake news induces abnormal trading and temporarily drives up the prices of small stocks, whose recent prior performance was deteriorating. These patterns are consistent with a motive to hire authors to write fake news to prop up the stock price" (Kogan et al., 2021).

Entertainment

Fake news can be created for humorous reasons or to entertain. This news also may be popular and attract the attention of users who may share this content with friends, comment, and get likes. This motive is different from the previous ones because a creator "does not have a certain aim as earning money or defaming some actors in the social or political sphere because the motive is the desire to make a joke to the audience" (Edson & Tandoc, 2019). People use the technologies simply for the fun of it or escapism. This phenomenon is quite widespread since user-generated content is omnipresent online, and the Internet allows everyone to create their own information on different topics. The appearance of the first "life journals" and blogs gave users the chance to be authors writing texts and communicating with others. Jenkins (2005) introduced such a concept as "participatory culture," which is a culture with relatively low barriers to artistic expression and civic engagement, and strong

support for creating and sharing one's creations. It means the process of connecting people on the Internet from different parts of the world, blurring the physical boundaries between countries, and giving people the unique opportunity to communicate and take part in the process of creating various elements to entertain users (comics, fanfic stories, etc.). At that time, people created this for fun and to express their creativity. At that time, despite the widespread UGC, the line between fantasy, unreal stories, and the real world was visible. Over time, the line between unreal and real worlds is blurred, and the reason for this is information overload. It has negative consequences because it leads to the appearance of different co-products as fake news. Creating fake news for entertainment motives is an example of when readers cannot understand the difference between real and unreal/fabricated elements. In turn, it might have a harmful effect: creators of fake news use the format of "news" as a format of information that a priori is considered reliable for mercenary purposes. This news pretends to be true, and it tends to convince people of its credibility and make them share it on the Internet. In this case, there is another hidden motive to mislead people that has a harmful effect and leads to "information pollution" and a decline in the quality of information in general.

Fear and panic

In this case, the desire to spread fear and panic among the public prevails. More often, it is typical for situations regarding health-related social issues, conflicts, wars, as well as natural disasters. It occurs due to the absence of verified information and the opportunity to check incoming information about events because of, mostly, physical distance and the inability to visit the scene of events. For this, social media are full of unverified content that may easily attract the attention of the public. On top, users are increasingly dependent on social media feed algorithms, which may create bubbles of news stories that only support personal beliefs. So, users may be misled upon reading the fake stories inside the bubble (Cheuk Hang Au, 2020). According to Kim et al. (2021), when epidemics become more common around the world, rumors of absurd and false medical tips spread rapidly on social media. When there is a lack of verified information, people feel great anxiety and fear and easily believe such tips, regardless of whether they are true (Hu et al, 2020; Jakovljevic et al, 2020). The period of COVID-19 also was noticed by spreading fake news about the virus and its negative consequences (Van Bavel et al., 2020; Pennycook et al., 2020). Another example is fake news about global warming, its reasons, and its possible impact in the future. It also gives rise to many conspiracy theories around environmental issues by presenting negative scenarios associated with natural disasters, water shortages soon, etc.

Status, attention, and identity building

Social needs might also be motives for producing fake news, such as status, attention, and identity building (see Wilson, 2020, Bente, 2021). Actors may create and share fake news to gain acceptance within online communities. Some scholars point out (e.g., Carey 1989; Bente, 2021) that communication is not the act of imparting information, but rather the representation of shared beliefs—communication draws people together in fellowship and commonality. In other words, writing fake news is a certain way of communication between a creator and the public. Creating fake news is a specific way to express yourself, draw attention to your personality, and a particular view of the world is portrayed and confirmed, and get feedback from the online community. By producing and sharing fake news (with a particular slant), users are connecting with other users. The second aim is to earn fame (Wardle & Derakhshan, 2017, p. 36). Social media users are incentivized through likes, shares, and comments to "create content that will resonate with their friends, followers, and groups," and media manipulation might be a way to gain status and express identity (Warwick & Lewis, 2017, p. 31). "Taken as a whole, these communities may feel that by manipulating media outlets, they gain some status and a measure of control over an entrenched and powerful institution, which many of them distrust and dislike" (Warwick & Lewis, 2017, p. 31).

4. Conclusion

This research has been conducted to generalize scientific knowledge on fake news, clarify different aspects regarding the features of its dissemination at different levels, and address some gaps. The main focus was on the motives of creators of fake news and the digital platforms and applications utilized in this process. The author provided a typology of actors (real human beings or non-humans), formats, and channels (social bots, cyborgs, trolls, fake news media, search engines and aggregators, deepfake news) through which fake news is disseminated. Second, the motives of the creators of fake news were shown, including political intentions, financial benefits, entertainment, fear and panic, status, attention, and identity building. Still, there is a need for further research. First of all, it is crucial to pay attention to the main reasons why people read, believe, and share

such news. To do this, it is necessary to conduct analyses on this topic at both theoretical and empirical levels. Additionally, another direction in which current academic knowledge is needed is conducting empirical research to gather relevant data about the influence of fake news on professional journalistic activity and the use of the latest approaches and applications for fact-checking.

References

Aldwairi, M., & Alwahedi, A. (2018) "Detecting Fake News in Social Media Networks", *Procedia Computer Science*, 141, pp. 215-222.

Allcott, H., & Gentzkow, M. (2017) "Social Media and Fake News in the 2016 Election", *Journal of Economic Perspectives*, 31(2), pp. 211-236.

Azzimonti M., Fernandes M. (2023) "Social media networks, fake news, and polarization", *European Journal of Political Economy,* Volume 76, https://doi.org/10.1016/j.ejpoleco.2022.102256.

Bakir, V., & McStay, A. (2017) "Fake news and the economy of emotions: problems, causes, solutions", *Digital Journalism*, 6(2), pp. 1-22.

Bente, K. (2021) "Fake News", Communication, [online], https://doi.org/10.1093/acrefore/9780190228613.013.809

Boshmaf, Y., Muslukhov, I., Beznosov, K., & Ripeanu, M. (2011) "The socialbot network: When bots socialize for fame and money", Proceedings of the 27th Annual Computer Security Applications Conference, pp. 93-102.

Carey, J. (1989) "*Communications as culture: Essays on media and society"*, New York: Routledge.

Cheuk Hang Au, W. (2018) "The Spread of Misinformation Online: Examining the Impact of Social Media and the Role of Automated Bots", *Journal of Media Studies*, 12(1), pp. 45-68.

Cheuk Hang Au, W. (2020) "Ideological Polarization and Fake News Consumption", *Journal of Communication Research*, 15(2), pp. 11-28.

Di Domenico, G., Sit, J., Ishizaka, A., & Nunan, D. (2021) "Fake news, social media, and marketing: A systematic review", *Journal of Business Research*, 124, pp. 329-341.

Edson C. Tandoc Jr., Zheng Wei Lim & Richard Ling (2018) "Defining "Fake News"", *Digital Journalism*, 6:2, pp. 137-153.

Edson C. Tandoc Jr. (2019) "The facts of fake news: A research review", *Sociology Compass*, 13(9), [online], https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12724

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016) "The rise of social bots", *Communications of the ACM*, 59(7), pp. 96-104.

Finkel J., Jiang S., Luo M., Mears R., Metaxa-Kakavouli D., …&Torres-Echeverry N. (2017) "Fake News and Misinformation", [online], https://pacscenter.stanford.edu/publication/fake-news-and-misinformation/

Gilbert, Ben (2016) "Fed up with fake news, Facebook users are solving the problem with a simple list", *Business Insider*, [online], https://acesse.one/spCa8

Hu, Z.; Yang, Z.; Li, Q.; Zhang, A.; Huang, Y. (2020) "Infodemiological Study on COVID-19 Epidemic and COVID-19 Infodemic", *Preprints*, [online], https://doi.org/10.20944/preprints202002.0380.v3

Jakovljevic M., Bjedov S., Jaksic N., Jakovljevic I. (2020) "COVID-19 pandemia and public and global mental health from the perspective of global health security", *Psychiatria Danubina*. 2020; p. 6–14. pmid:32303023

Jenkins, H. (2005) "*Convergence culture: Where old and new media collide*", New York: NYU Press.

Kasra, M. (2017) "Fake news: The role of media and bias in shaping public opinion", *Media Studies Journal*, 21(1), pp.14-27.

Kim, J., Moravec, P. L., & Dennis, A. R. (2021) "Combating fake news on social media with source ratings: The effects of user and expert reputation ratings", *Journal of Management Information Systems*, 38(3), pp. 1018-1047.

Kogan, N., Gaskins, B., & Goss, M. (2021) "The impact of fake news on stock markets: A pump-and-dump scheme", *Journal of Financial Markets*, 10(2), pp. 89-106.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018) "The science of fake news", *Science*, 359(6380), 1094-1096.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., Nguyen, C. M. (2022) "Deep learning for deepfakes creation and detection: A survey", *Computer Vision and Image Understanding,* Volume 223, 2022, 103525, ISSN 1077-3142, https://doi.org/10.1016/j.cviu.2022.103525.

Olmstead K., Mitchell A., Rosenstiel T. (2011) "Navigating News Online: Where People Go, How They Get There and What Lures Them Away", Pew Research Center's Project for Excellence in Journalism, [online], https://www.pewresearch.org/wp-content/uploads/sites/8/legacy/NIELSEN-STUDY-Copy.pdf> accessed 30 October 2021.

Pennycook G., McPhetres J., Zhang Y., Lu J., Rand D. (2020) "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention", *Psychological Science*, p. 770–780.

Rocha, Y. M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., & de Figueiredo Nicolete, L. D. (2021) "The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review", *Z Gesundh Wiss*, pp. 1-10.

Rochlin, N. (2017) "Fake news: belief in post-truth", *Library Hi Tech*, 35(3), pp. 386-392.

Shao, C., Ciampaglia, G.L., Varol, O. et al. (2018) "The spread of low-credibility content by social bots", *Nat Commun 9*, 4787, [online], https://doi.org/10.1038/s41467-018-06930-7

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017) "Fake News Detection on Social Media: A Data Mining Perspective", *ACM SIGKDD Explorations Newsletter*, 19(1), pp. 22-36.

Tandoc, E. C., & Albright, J. (2016) "All the news that's fit to click: The economics of clickbait media", *Journalism Practice*, 10(6), pp. 836-848.

Townsend, L. (2017) "The Impact of Fake News on Public Opinion", *Media Studies Review*, 9(3), pp. 87-102.

Van Bavel JJ., Baicker K., Boggio P., Capraro V., Cichocka A., Cikara M., et al. (2020) "Using social and behavioural science to support COVID-19 pandemic response", *Nature Human Behaviour*, p. 460–471.

Wardle, C., & Derakhshan, H. (2017) "*Information disorder: Toward an interdisciplinary framework for research and policy making*", Strasbourg: Council of Europe report DGI (2017)09.

Warwick, A., & Lewis, R. (2017) "*Media manipulation and disinformation online"*, New York: Data & Society.

Weir, C. (2009) "Understanding the Social Dynamics of Fake News: A Study of Online Behavior", *Journal of Information Technology*, 23(4), pp. 299-310.

Wilson, B. (2022) "The impact of fake news on media consumption behavior", *Journal of Digital Marketing*, p. 1-22.

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019) "The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans", *Journal of Data and Information Quality (JDIQ)*, 11(3), pp. 1-37.

Zhang, X., & Ghorbani, A. A. (2018) "Fake News Detection: A Data Mining Perspective", *Elsevier*, [online], https://www.sciencedirect.com/science/article/abs/pii/S0306457318306794

Towards the Virtual: AI Beauty Pageants and Their Implications for Society

Ida Knapik
AGH University, Kraków, Poland

ikna@agh.edu.pl

Abstract: The dynamic growth of Artificial Intelligence is accelerating rapidly, making AI present in almost every aspect of human life. Digital assistants available on smartphones, smart homes, autonomous cars, and even recently launched AI beauty pageants. The very first Miss AI pageant, announced by the World AI Creator Awards (WAICA) program and Fanvue company, marks a new era in beauty contests. The main idea of Miss AI is to present an AI-based character, which is evaluated based on its beauty, poise, and response to typical beauty pageant questions. What makes it innovative is that Artificial Intelligence not only participates but also serves as a judge. This is because the jury, in addition to normal humans, includes virtual influencers, like Spanish digital mouse Aitana Lopez. Beauty contests have always been controversial - after all, they involve judging individuals by their appearance and setting beauty standards. However, including Artificial Intelligence in this phenomenon raises even more profound ethical questions and blurs the boundary between the virtual world and the real world. This article focuses on researching how the idea of Miss AI impacts its audience, who primarily belong to a social media-driven society. The author used content analysis and case study methods to examine the topic. Drawing on the example of one of the most popular virtual influencers, Aitana Lopez, and her increasing popularity on social media, especially on Instagram, the author reflects on the ethical implications of the Miss AI concept and its ongoing perception by society. Due to the continuous growth of AI beauty pageants, the article provides initial insights into the topic and suggests potential future developments. What can be said for certain is that recent years have brought a significant AI revolution that affects and will continue to affect a growing number of dimensions of human existence. However, this raises many concerns, increasingly centered around ethics.

Keywords: Artificial intelligence (AI), AI beauty pageant, Virtual influencer, Ethical implications

1. Introduction

The first modern beauty contests appeared in the 1920s, starting a new era of setting beauty canons. During their early years, the most significant international beauty pageant was *Miss Universe*, held in the USA (Mihăilă, 2017). Over 100 years later, a new variant of international beauty contests is emerging. Artificial Intelligence (AI) has become a part of a debatable topic - setting beauty standards. During the last few years, a *technological tsunami sweeping the globe* has been observed (Elliott, 2019). Multiple advancements in AI topic are the consequences of the Fourth Industrial Revolution (Darko et al., 2020). Various researchers have already studied the sentiment towards AI and its social and ethical implications. Bao et al. (2022) have classified AI perceptions into five segments - negative, ambivalent, tepid, ambiguous, and indifferent classes. The researchers have proven that perspectives on AI differ based on the amount of news coverage and the content that audiences engage with. Floridi and Cowls (2022) proposed a framework consisting of five principles for AI ethics, emphasizing that the development and use of AI can both positively and negatively impact society. AI can reduce or intensify existing inequalities, solve longstanding problems, or create new ones. Neudert et al. (2020) demonstrated that many people are concerned about the risks of using AI after conducting extensive studies across 142 nations. AI beauty pageants, being a new phenomenon, have not yet been studied in terms of societal and ethical implications. This led the author to fill this gap and pose the following research question - *How does social media society perceive AI beauty pageants?* Additionally, the author hypothesized that *AI beauty pageants are negatively perceived by social media society.* Apart from the introduction, the paper consists of four sections aiming to answer the underlying research question.

1.1 Introduction to AI Beauty Pageants

Miss AI, being the first AI beauty pageant, is organized by Fanvue company in collaboration with the World AI Creator Awards (WAICAs) program, whose goal is to recognize the achievements of developers using AI worldwide. The main idea of the contest is to create an AI-based model with immense popularity on social media, which is one of the components of earning the Miss AI crown. The jury consists of two real human beings – Andrew Bloch, media founder and entrepreneur, and Sally-Ann Fawcett, beauty pageant historian, as well as two AI-based virtual influencers – Emily Pellegrini and Aitana Lopez. Contestants are judged based on traditional pageantry criteria, including beauty, poise, and unique responses to questions like *if you could have one dream to make the world a better place, what would it be?,* the technology used to create them, and social media popularity. The total prize pool for the competition is $20k. At the time of writing this paper, contest entries are officially closed, and 10 finalists across the globe, including Portugal, France, Brazil or India, have been selected.

The award ceremony date is still unknown. Considering the emerging trend of AI beauty pageants, the author aimed to gather initial opinions and insights based on existing comments.

2. Research Methodology

For the purpose of investigating the formulated research question, the author used a two-prolonged approach, which consisted of content analysis and case study methods. Initially, data obtained from users' comments available on Instagram for analysis purposes has been gathered. In total, 475 comments have been collected manually from 11 different accounts (number of comments as of 1 June 2024). Two of the accounts are run by human juries, Sally-Ann Fawcett and Andrew Bloch, while the rest of the accounts are news-related. The reason why the author decided to review the comments manually is because some of the comments had to be excluded, for example, by tagging other users or off-topic content. Emojis, which appeared frequently, had to be excluded from the initial analysis in the tool but included in the manual analysis. The rationale for the author's decision not to replace the graphic forms of the emoticons with their textual name in the tool analysis is that it could distort the results in this tool, blurring the distinction between the emoji name and the actual comment. GIFs have been completely omitted in both analyses. Figure 1 represents graphically how many comments have been collected from each account. The author reviewed Instagram by searching *miss AI* and *AI beauty pageant(s)* hashtags to find the most recent and the most commented posts across the platform.

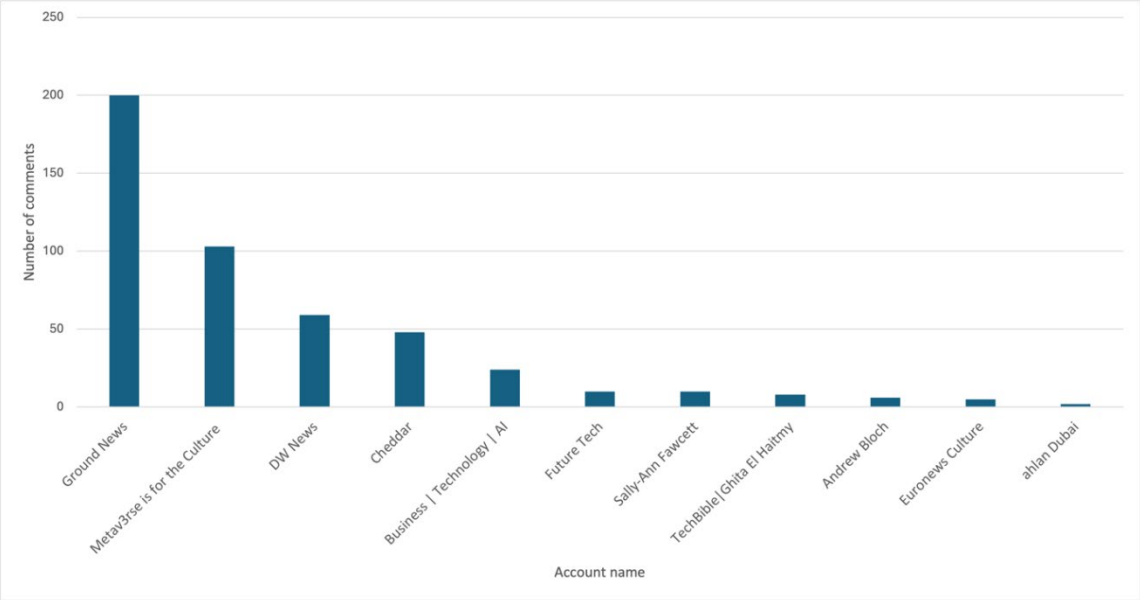


Figure 1: Number of comments collected from each account

In the second phase, comments analysis in R language have been conducted using the RStudio tool. A CSV file has been created and loaded into the application. Retrieved data has been properly transformed with the aim of creating a word cloud and performing sentiment analysis. Because the analysis in the tool has been conducted based on individual words, which in many cases changed the original tone of the commentary, as a third part, the author decided to review the comments manually and include emojis. Human-coder analysis (Alafwan et al., 2023) has been executed on 475 comments, putting the emphasis on the virtual jury, Aitana Lopez, who was positioned as a central figure associated with AI beauty pageant posts. The author incorporated mixed approaches in comments analysis to investigate users' initial opinions and aimed to identify possible future trends in the phenomenon of AI beauty pageants, comparing both single words and entire comments.

3. Results

This study used a mixed-methods approach, including both quantitative and qualitative data collection and analysis. The results are divided into two parts - the first one obtained using the designated tool and the second one derived from a manual review. In both cases, gathered users' comments have been analyzed and compared with one another to draw proper conclusions. Together, these results provide a comprehensive understanding of preliminary opinions on AI beauty pageants across social media society.

3.1 Single Words Analysis Output

The initial state of all words, including common English stop words (e.g. *the*, *to*, *a, and*) in total, contained 1549 words occurring with varying frequency. To create a word cloud, the dataset was cleaned by removing stop words, numbers, and punctuation, which are unnecessary for analysis in the tool. As a result of the comments review conducted using the RStudio environment, the following word cloud has been obtained (Figure 2). The author has chosen to present the most frequent words which appeared in comments at least five times. The top 10 most frequently words are as follows: *can* (37 times), *women* (37 times), *beauty* (32 times), *people* (28 times), *pageant* (25 times), *like* (24 times), *now* (21 times), *real* (19 times), *world* (17 times) and *human* (14 times). Among less common words, however, deserving attention in the AI beauty pageants context, the following words have been collected: *contest* (10 times), *create* (10 times), *standards* (10 times), *art* (8 times), *girls* (8 times), *models* (8 times), *porn* (8 times), *sex* (8 times), and *fake* (6 times).



Figure 2: Word cloud representing most frequently used words, which appeared at least 5 times

A sentiment analysis has been conducted to determine how words from the comments are categorized in terms of feelings. To do that, a proper sentiment lexicon had to be selected. A sentiment lexicon is a collection of words and expressions, each given a positive or negative rating, which indicates its sentiment polarity and strength (Tang et al., 2014). After reviewing available options, the author decided to use the Bing sentiment lexicon (Hu & Liu, 2004), which classifies words in a binary fashion, either positive or negative. Applying the lexicon reduced the studied dataset to 215 words. There is also another reason behind choosing that lexicon and applying bipolar split. Studied data contained too few observations to draw more tendencies from it, and as stated in the Introduction section, AI beauty pageant is a new phenomenon and is at its beginnings of growth. Figure 3 shows the top 10 most frequent words categorized by the sentiment. In the negatives, the most common word is *miss.* As stated in Cambridge Dictionary (2024), *miss* is both a verb and a noun and has multiple meanings. Therefore, having this word categorized as a negative means that it has been taken as a verb or noun meaning a failure, not a girl or a woman, which should have been the case here. The most frequent word in this analysis and included in the Bing lexicon is *beauty*, classified as a positive.

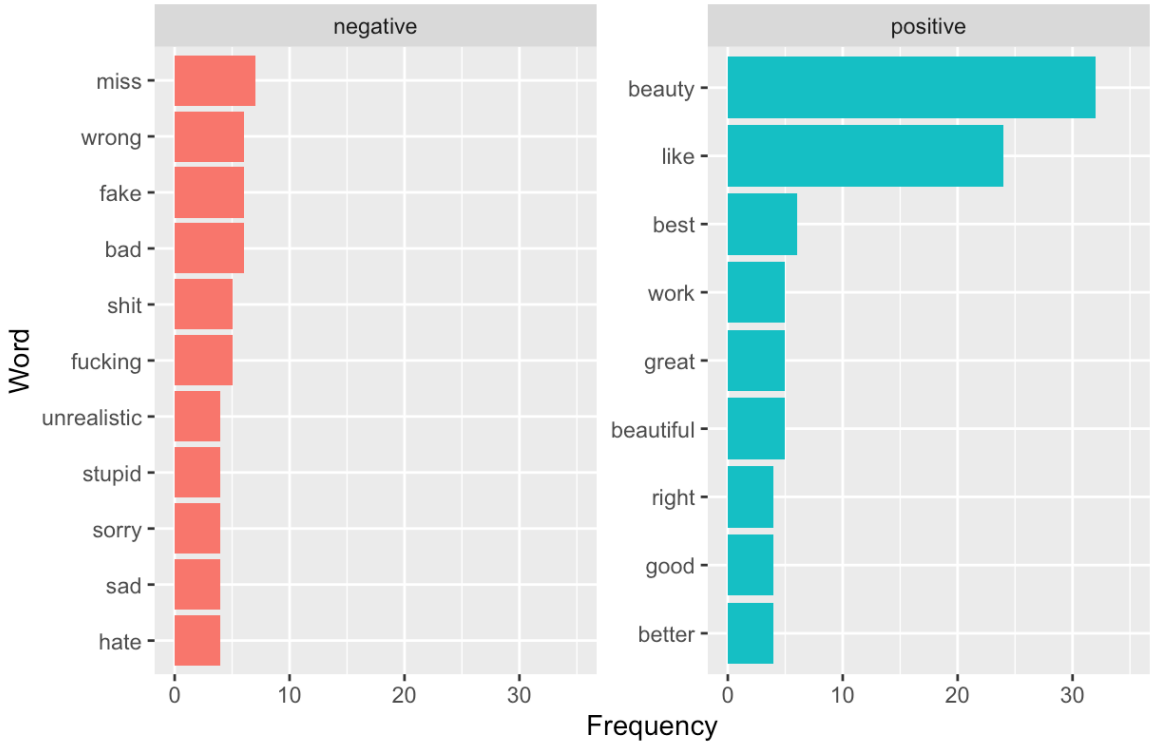


Figure 3: Top 10 most frequent words classified by sentiment (positive and negative)

Considering this result, the studied dataset expresses both positive and negative sentiments, with an advantage on the positive side. This is because *beauty* and *like* terms, one of the most frequent words available in the comments, are classified per the Bing lexicon as positive. The author has also chosen to examine the sentiment distribution encompassing all analyzed words. As presented in Figure 4, more than a half, specifically 54%, which are 116 items, were categorized as negative, and less than a half, more precisely 46%, so 99 words as positive.

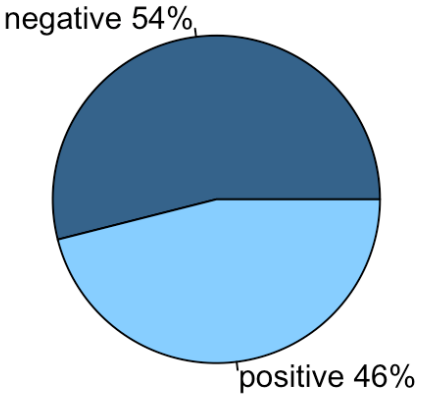


Figure 4: The overall sentiment distribution across all studied words

Given this output, it can be detected that the studied group of people is more likely to perceive AI beauty pageants negatively. However, sentiment is split almost half by half, so the general feelings are not unambiguous yet. Nevertheless, the analysis was performed based on single words, which in many cases doesn't reflect the original tone and misses human understanding. This is why the author decided to analyze whole comments manually, which is described in the next section.

3.2 Comments Analysis Output

To reveal a wider context of the studied dataset, as the second part of the research, the author performed a manual comments analysis. To demonstrate it, 12 examples of comments containing words classified according

to the Bing lexicon were selected, carefully read, and assessed by a human (the author). Each comment has been copied and pasted exactly as originally written on Instagram, including all punctuation and typographical errors. Table 1 gathers the comments, listing the lexicon's words and the author's opinion on the general feeling of the entire comment, without splitting it word by word. Furthermore, the manual analysis included emojis (if any), which the tool struggles to analyze. Due to possible errors in the graphical display of emojis in the article, the author decided to write the name of it in parentheses.

Table 1: Studied comments and lexicon words with the author's opinion

Comment	Words from Bing lexicon and its sentiment	Author's sentiment
Fake beauty standards **(unamused face)**	*fake* (negative) *beauty* (positive)	negative
This is what we are doing with AI, not making excel better **(angry face)** ***what a waste of*** **(man facepalming)**	*excel* (positive) *better* (positive) *waste* (negative)	negative
Sorry… but this is sick! Just thinking about young girls and how this is affecting their beauty standards. **(pensive face)**	*sorry* (negative) *sick* (negative) *beauty* (positive)	negative
Women are complaining about beauty standards now, just imagine what's gonna happen in 10 years when boys are creating their dream girls on a computer with unmatched beauty	*complaining* (negative) *beauty* (positive) *unmatched* (positive)	negative
This is so toxic, and wrong in a lot of levels.	*toxic (negative)* *wrong (negative)*	negative
a sad day for humanity and the beautiful and real ladies around the world...Perfection does not exist...but yes real people with perfect imperfections....	*sad* (negative) *beautiful* (positive) *perfect* (positive) *perfection* (positive)	negative
Although how is AI good or great but it is still not human **(robot face)**	*good* (positive) *great* (positive)	negative
at least pervs will have a "good looking woman" online **(face with rolling eyes)**	*good* (positive)	negative
Lol. Late Stage Capitalism is having AI beauty contests. Pathetic	*beauty* (positive) *pathetic* (negative)	negative
Great, so now we the women will be subjected no to just human "beautiful" standards, now will be artificial standards…	*great* (positive) *beautiful* (positive)	negative
Remember when AI was suppose to make life easier. **(thinking face)**	*easier* (positive)	negative
This is disgusting **(face vomiting)**	*disgusting* (negative)	negative

Based on the above comparison can be observed that the results of the single-word analysis slightly differ from the results of the whole comments analysis. Although some comments contained positively classified words, the overall sentiment of the sentence conveyed a negative feeling. In many cases, the positively classified word was *beauty* included in the main topic name *AI beauty pageant*. This is why it may disrupt the outcome of the overall sentiment, increasing the positives. However, the manual review corrected these errors and validated the findings, as shown in the examples in Table 1. What's more, supplementing the tool method with manual analysis incorporated human intuition and understanding, which are essential for accurately interpreting complex emotions and sentiments. Given the outcome of the human analysis, the majority of the comments expressed a negative sentiment, which means that the perception of the AI beauty pageant among social media

society is pessimistic. Moreover, while studying the dataset and embedded impressions, the author observed a few main directions that appeared in the comments. Table 2 gathers 14 distinguished topics, which recurred among users' opinions.

Table 2: Main topics observed in comments

Theme	Sample comments
Surprise and disbelief	*Wow, I wasn't expecting AI to be so influential in every sector so soon. Seems like in less than a year, everywhere you look.*
Negative impact on young girls	*This is extremely dangerous and damaging to young girls trying to meet beauty standards of nonexistent images; Imagine a situation a girl looking at the mirror flaunting her beauty at the same time one pop up just appear on phone and laptop for a sec with the line "Am more beautiful then you!!!!" and just vanish…*
Concerns about future implications	*This is all going in the wrong direction; Stop the madness, this is the wrong direction.*
Objectification of women and concerns about sexism	*Women are complaining about beauty standards now, just imagine what is gonna happen in 10 years when boys are creating their dream girls on a computer with unmatched beauty; Men reinventing the objectification of women.*
Comparisons to human beauty pageants	*Beauty contests full stop should be banned; As if real beauty pageants weren't bad enough, now it is create your own.*
Misuse of technology	*Imagine if this energy was put towards something actually meaningful and productive; What a stupid use of technology*
Sarcasm and jokes	*For the talent portion they just see who can solve a Facebook login captcha*
AI as art	*No because ai isn't art; That is subjective, just like art tbh*
Collapse of society	*We are witnessing the collapse of our society; We're cooked as a society.*
Money laundering	*New money laundering playgrounds?; Money laundering again.*
Late-stage Capitalism	*Lol. Late Stage Capitalism is having AI beauty contests. Pathetic; Late stage capitalism at its finest*
Pornography	*Nobody is surprised by this. The internet was created and men used it for porn. AI will be no different, unfortunately; Poor people use it for porn and will continue to use everything for porn.*
Consequences of Industrial Revolution	*The Industrial Revolution and its consequences…*
Artificial beauty standards	*Great, so now we the women will be subjected no to just human "beautiful" standards, now will be artificial standards…*
Cultural implications	*The West has fallen*

As can be clearly observed, the dataset reflects a predominantly mix of negative sentiments towards the idea of an AI beauty pageant, with significant concerns about its impact on society, particularly regarding beauty standards, objectification of women, and pornography. While some humour is expressed, the overarching response is critical, highlighting ethical and societal issues and implications. In the context of comments, graphical posts should be mentioned as well. One of the most frequently posted characters in the AI beauty pageant range is Aitana Lopez. Users have not missed pointing this out, adding a proper note to the discussion. This leads to a short case study of Aitana, based on the users' comments and general perception of the character.

3.2.1 Aitana Lopez case study

Aitana Lopez, one of the jury members of the AI beauty pageant, has been created by The Clueless agency, based in Barcelona, Spain (2024). As of now (22 June 2024), she has more than 321k followers on Instagram, and this number is still increasing. It can be assumed that being a jury in the contest and her immense popularity were leveraged in social media posts announcing the AI beauty pageant. Therefore, while reviewing all the comments,

the author noticed that some of the sentences refer precisely to Aitana, not to the AI beauty pageant itself. Instagram users primarily paid attention to Aitana's appearance and, more specifically, which famous person she resembles. Among the similarities to the celebrities have been observed Lady Gaga, Nicole Scherzinger (*This is absolutely Gaga mixed with Nicole Scherzinger*), Gal Gadot (*That's just Gal Gadot with pink hair*), Dua Lipa (*Isn't that just dua lipa?*), and Jessica Alba (*Didn't they already do this years ago and it kinda looked like Jessica Alba?*). There has also been specified a second group of comments on Aitana's look - *She looks great*; *She looks good; Pink hair looks like a dude with a wig* or *Love you sis.* Based on that, it can be observed that people present mixed feelings on this topic. On the one hand, they find Aitana like a celebrity and compliment her look, but on the other hand, they express themselves negatively about her. Considering all the results, it can be recognized that this case study also demonstrates a bipolarity of sentiments.

Summarizing all the analyses, both performed in the tool and the manual one, it can be said that social media society perceives AI beauty pageants in a negative light. While the tool has shown the sentiment is spread almost fifty-fifty, the qualitative overview was required to complement the results through a holistic interpretation of the data. The chosen methods allowed for the application of a mixed and balanced approach, which brought comprehensive insights, increased the reliability of the results and granted a nuanced understanding of the sentiment in the comments.

4. Discussion

The article presents a study on social media society's perception of a new AI phenomenon, which is AI beauty pageants. The study has revealed a few key findings. First, the respondents' overall attitude towards the main topic is predominantly negative, with significant criticism, scepticism, and numberless concerns. There are loads of opinions acknowledging real beauty canons, which might be dominated by unrealistic and harmful beauty standards and a fear of a dystopian society where human values are flawed by artificial standards. Furthermore, the objectification aspect, criticizing how pageants might contribute to the disruption of women's perception, marking AI beauty pageants as a male-dominated initiative. One of the previous studies about public perception of AI has shown that more than half of the surveyed think that AI will lead to the downfall of society in different contexts, *inter alia*, the personal, the social, and the cultural. Furthermore, the researchers have proven that AI and its implications were not seen in simple black-and-white terms. The participants held nuanced views on how AI will impact their lives (Brauner et al., 2023). The study described in this article points out various aspects and nuances in terms of AI usage, similar to Braunder's studies, but in a beauty pageant context. Most of them express negative attitudes, however, there is a wide range of raised issues. It is essential to note that this study was conducted on a small dataset, and the used tool is not limitations-free, e.g. by a limited number of records in the lexicon. In another study, Araujo et al. (2020) explored the utility of AI, *inter alia*, in media. Research participants expressed common concerns about the risks associated with AI and questioned its fairness and usefulness for society. According to scientists, to ensure the appropriate adoption of AI, it is crucial to acknowledge end-user perceptions and risk assessments at both individual and societal levels. Considering the results shown in this article, end-users' perspectives and possible implications have not been sufficiently analysed and assessed by the originators. At the current early stage of the AI contest, there is a clear reluctance on both grades, personal and social. However, it is important to note that this study was conducted based on a relatively small group of people and comments, as the AI beauty pageant is a new circumstance and will evolve with time. In connection with implications, the Collingridge dilemma should be quoted. The dilemma states that if the technology is sufficiently advanced and accessible, it can be thoroughly evaluated, however, usually, it is too late to regulate its development. On the contrary, if the technology is new and not yet widespread, assessing its future perception and likely impact is challenging, but managing its development and usage is easier (Collingridge, 1982). Responsible research and innovation somehow obligate researchers to continuous updates of the understanding of societal evaluations and implications of technology development (Burget et al., 2017; Owen & Pansera, 2019). The dilemma documented many years ago is versatile and finds its use even in AI beauty pageants. It is hard to predict how the AI contests will impact society in the future, not only social media but people in general. Based on the conducted study, the first directions have been identified, however, the forthcoming is quite uncertain and may bring unexpected consequences. Nevertheless, having disclosed initial insights, the development and use could be improved in order to prevent the increasingly unfavorable implications. This leads to another significant finding, which is ethical implications. According to the publication by OECD (2019), AI could cause ethical issues and social concerns. The main statements revealed in this article present that the boundary between the virtual and the real world is permeating. Starting with the creation of idealized female forms designed to cater to specific aesthetic preferences, through the creation of hyper-realistic and idealized images, which could dehumanize relationships and interactions, leading to the superiority of

artificial companions over real human connections. Next is the alteration of expectations of beauty and relationships, making it harder for real individuals to meet new, unrealistic canons. From a psychological perspective, AI beauty standards may contribute to mental health issues, including anxiety and depression. Finally, the exclusion of normal people from being appreciated for their natural appearance, marginalizing those who do not meet the AI norms. Totalling the quoted implications, one cannot disagree that normative and regulatory frameworks of AI, especially ethics are crucial (Cannarsa, 2021; Brauner et al., 2023). This study has revealed the tip of the iceberg, which is still growing.

5. Conclusions

Unquestionably, the growing presence of AI solutions in nearly every aspect of human life is transforming and will transform many areas. AI influence extends beyond jobs and daily life tasks, now permeating the beauty contests domain. As indicated in this article, there are new circumstances which so far impact social media society. AI beauty pageants, being little explored yet at the beginning of their existence, do not meet with a positive reception by the public. Conversely, people express their concerns and the negative implications brought by this new phenomenon. In explicit opinions, a few main themes have been identified and described in this paper. As stated by Johannessen (2020), embracing new technologies implies the acquisition of new skill sets, particularly those enabling ethical reflection on the complex issues arising from such advancements. In the instance of AI beauty pageants and their early but already observed negative social and ethical implications, people will be forced to learn how to handle it. Certainly, by observing and documenting initial observations on the topic in this writing, society can proactively address issues and mitigate potential negative long-term consequences. However, looking at the amount of negative feelings towards the subject matter can be challenging. As further steps of research on the AI beauty pageant topic, the author suggests studying a wider range of data (comments), which will gradually appear. Additionally, based on the experience gained in this research, gathered comments could be investigated in a more complex tool, which can read the graphics (emojis, GIFs). Next, results could be compared with results from this paper and subjected to human assessment. Will AI settle human canons of beauty? Hopefully not, as that could mark a slow demise of humankind.

References

Alafwan, B., Manahan Siallagan and Utomo Sarjono Putro (2023). Comments Analysis on Social Media: A Review. *ICST Transactions on Scalable Information Systems*. doi:https://doi.org/10.4108/eetsis.3843.

Araujo, T., Helberger, N., Kruikemeier, S. and de Vreese, C.H. (2020). In AI We trust? Perceptions about Automated decision-making by Artificial Intelligence. *AI & SOCIETY*, 35(3). doi:https://doi.org/10.1007/s00146-019-00931-w.

Brauner, P., Hick, A., Philipsen, R. and Ziefle, M. (2023). What does the public think about artificial intelligence?—A criticality map to understand bias in the public perception of AI. *Frontiers in Computer Science, 5.* Available online at: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2023.1113903/full (Accessed 22 June 2024).

Burget, M., Bardone, E. and Pedaste, M. (2016). Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review. Science and Engineering Ethics, 23(1), pp.1–19. doi:https://doi.org/10.1007/s11948-016-9782-1.

Cambridge Dictionary (2024). *miss*. @CambridgeWords. Available online at: https://dictionary.cambridge.org/dictionary/english/miss (Accessed 15 June 2024).

Cannarsa, M. (2021). Ethics guidelines for trustworthy AI. *The Cambridge handbook of lawyering in the digital age*, 283-97.

Collingridge, D. (1982). *Social Control of Technology*. Continuum International Publishing Group Ltd.

Darko, A., Chan, A. P., Adabre, M. A., Edwards, D. J., Hosseini, M. R., & Ameyaw, E. E. (2020). Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. *Automation in construction*, 112, 103081.

Elliott, A. (2019). *The culture of AI: Everyday life and the digital revolution*. Routledge.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, pp.168–177. doi:https://doi.org/10.1145/1014052.1014073.

Johannessen, J. A. (2020). *Artificial intelligence, automation and the future of competence at work.* Routledge.

Mihăilă, V. (2017) 'The Making of Modern Beauty Pageants Traditions, Myth, Tales, and Histories', *Yearbook of the 'Gheorghe Sincai' Institute for Social Sciences & the Humanities of the Romanian Academy,* 20, pp. 107–125. Available online at: https://search-1ebscohost-1com-1000003lc0080.wbg2.bg.agh.edu.pl/login.aspx?direct=true&db=asn&AN=127864412&lang=pl&site=ehost-live (Accessed 28 June 2024).

Neudert, L. M., Knuutila, A., & Howard, P. N. (2020). Global attitudes towards AI, machine learning & automated decision making. *Google Scholar Google Scholar Reference.*

OECD. 2019. *Artificial Intelligence in Society*. Organisation for Economic Co-Operation and Development Publishing. Available online at: https://www.oecd-ilibrary.org/sites/eedfee77-en/index.html?itemId=/content/publication/eedfee77-en (Accessed 26 June 2024).

Owen, R., and Pansera, M. (2019). Responsible innovation and responsible research and innovation. *Handbook Sci. Public Policy*, pp.26–48. doi: 10.4337/9781784715946.00010.

Tang, D., Wei, F., Qin, B., Zhou, M. and Liu, T. (2014). *Building Large-Scale Twitter-Specific Sentiment Lexicon : A Representation Learning Approach*, pp.172–182. Available online at: https://aclanthology.org/C14-1018.pdf (Accessed 20 June 2024).

www.theclueless.ai. (2023). *Aitana López - The Clueless.* Available online at: https://www.theclueless.ai/project/aitana-lopez. (Accessed 22 June 2024).

Masters Research Papers

Exploring Student Perspectives on Generative AI in Requirements Engineering Education

Nicklas Mellqvist and Peter Mozelius
Mid Sweden University, Östersund, Sweden

Nicklas.Mellquist@miun.se
Peter.Mozelius@miun.se

Abstract: The rapid development of generative AI (GenAI) technologies in recent years has enabled new opportunities as well as new challenges in higher education. While many studies in computer science have focused on GenAI in programming education, fewer have examined its possibilities and challenges in requirements engineering (RE). This study aims to explore the impact of GenAI on the pedagogical aspects of RE in higher education, focusing on the student perspective, to analyse how GenAI might influence learning experiences, knowledge acquisition, and skill development. The main research question to answer was: "What are the students' perspectives of the integration of GenAI in the educational practices of requirements engineering?" An Action research strategy was employed, with one of the authors also serving as teacher in the investigated course. A mixed-methods approach was used to collect both qualitative and quantitative data from workshops and surveys. During the workshops, students used ChatGPT to generate and evaluate software requirements and compared these to manually crafted requirements. Thematic analysis of the qualitative data captured students' perspectives, while survey data identified trends and preferences. Findings show that while students generally had a positive experience with GenAI, valuing its efficiency and the quality of generated requirements, they also recognized the need for human oversight to maintain accuracy. The study highlights both opportunities and challenges of using GenAI in RE education. While GenAI increased learning engagement and helped with brainstorming, students faced difficulties in creating effective prompts and found it time-consuming to refine AI-generated requirements. A hybrid approach, combining AI-generated and manually created requirements, proved most effective by balancing AI's advantages with human insights. Further research is needed on how GenAI could be effectively integrated into computer science education.

Keywords: Generative AI, Requirements engineering, Mixed methods, Higher education, Action research

1. Introduction

There has been a fast development of tools that build around Generative AI (GenAI) techniques after the release of ChatGPT in 2022. This rapid development of different GenAI tools has created both promising opportunities as well as new reports on challenges in higher education (Chan and Hu, 2023; Yeralan and Lee, 2023). In the field of computer science, there are studies reporting that students' use of GenAI in programming education can be problematic when syntactically correct code is generated without any deeper understanding of code semantics (Denny et al, 2024). Programming courses in higher education should also develop critical thinking, which might be lost in a frequent use of GenAI-based code construction (Kasneci et al., 2023).

Many research studies can be found that investigate the challenges and opportunities of GenAI in programming education. Less studies have been conducted about the possibilities and challenges of involving GenAI in requirements engineering (RE), and most of them have investigated RE in the industry and not in academia. A recent study by Ronanki et al. (2024), highlight that prompt engineering is a critical factor for success when using GenAI for RE. At the same time, this study also brings up the lack of tools and methods to systematically assess and create the effective prompt patterns to employ for specific RE tasks. Despite the advancements and potential opportunities, the use of GenAI in RE education remains underexplored. A literature search revealed that most studies focus broadly on AI in education rather than on GenAI tools specifically within RE education settings. This reveals a research gap and highlights a need to investigate the impact and optimal practices for integrating GenAI in this educational context. However, several challenges with integrating GenAI into higher education have been identified. For example, there are concerns regarding academic integrity and the responsible use of GenAI in educational settings, which requires careful supervision (Yu, 2023; Sullivan et al., 2023). This study explored both the perceived affordances with GenAI in requirements engineering, and the aligned constraints in the described research area.

The aim of this study is to explore the impact of GenAI on the pedagogical aspects of higher education RE. Moreover, the focus of this study was on the student perspective, and to analyse and discuss how GenAI might influence students' learning experiences, knowledge acquisition, and skill development. The research question that guided this study was: "What are the students' perspectives of the integration of GenAI in the educational practices of requirements engineering?"

2. Extended Background

Requirements engineering (RE) is generally accepted as critical to the creation of high-quality software (Ochodek, 2023). Engineering Education (REE) must teach students to specify, elicit, and manage requirements from stakeholders while also addressing socio-technical challenges and human-related aspects, preparing them for industrial challenges (Daun et al., 2022; Spichkova M, 2019). Recent trends in RE education (REE) show a shifting focus from instructor-centric approaches towards industry-centric approaches including collaborative project-based settings and the involvement of external stakeholders, which aims to give students an authentic RE experience (Daun et al., 2022).

GenAI has the potential to enhance learning experiences in higher education and is likely to drive innovation and improvement in the coming years, while also creating several new challenges (Michel Villarreal et al., 2023). In REE, AI opens up new opportunities for teaching and learning, and can assist students with practical RE tasks, ranging from brainstorming to requirements elicitation and validation (Moravánszky, 2024). GenAI tools can potentially contribute to making education more engaging and immersive due to its dynamic and interactive nature (Michel Villarreal et al., 2023). Its ability to understand and generate content similar to what humans produce allows providing personalized feedback to students (Daun and Brings, 2023). This can improve student engagement, motivation and retention, due to its ability to adapt to different students' needs and abilities (Hartley, 2023).

Uncontrolled student engagement with AI tools can lead to students relying too much on the AI-generated material, which can potentially diminish their critical thinking and problem-solving skills (Kasneci et al., 2023). To avoid the risk of GenAI tools "doing the homework" for learners and simultaneously boosting their understanding, Petrovska et al. (2024) suggests using AI tools in a careful and creative way by developing assessments that can encourage learners to critically evaluate AI-generated content. In a recent REE study by Moravánszky (2024), challenges were found regarding creating effective prompts, and AI was valued as a supplementary tool rather than a complete replacement. This emphasizes AI's optimal performance when combined with human expertise and guidance. It is important to strike a balance between new technology and student-centred approaches to ensure that the integration of GenAI improves education rather than detracts from it. Human skills such as critical thinking, curiosity and personal growth cannot be replaced by an AI (Shaji George, 2023).

3. Methodology

This chapter presents the methodology and materials used in this study. It contains descriptions of the methods of research, data collections, and data analysis.

3.1 Action Research Approach

This study was conducted according to the action research approach outlined by Norton (2009). In the spirit of action research one of the authors has been a teacher in the investigated course and moderating the workshops where data were collected. Action research has been defined as a practically oriented, but systematic research approach, "*that enables teachers to investigate their own teaching and their student's learning"* (Nolen & Putten, 2007, p. 401). There are several branches of Action research, where this study should be classified as in the British tradition that has been described by Norton (2009, p. 71), as an approach *"that links research to improvement of practice and is education orientated"*. The authors' overall objective is to achieve research results that can be useful in the future redesign of activities in the investigated course.

3.2 Data Collection

A mixed methods approach was used, combining both qualitative and quantitative methods, to provide a comprehensive understanding of how students interact with GenAI in RE education. This approach enriches the analysis where the combination of different data sources strengthens the validity (Creswell & Plano Clark, 2017). An exploratory sequential design was employed, starting with collecting and analysing qualitative data to help develop the quantitative survey. The process began with workshops that included reflections and group discussions to gather detailed insights. These insights were then used to develop a survey. This approach ensured that the survey was based on real student experiences, making it more relevant and accurate (Creswell & Plano Clark, 2017).

3.2.1 Workshops

The workshops were an integral part of the research design. They were conducted during the course "Analysis and Design of Information Systems", which focuses on requirements management and design principles. Each workshop included the following components:

- **Interactive Sessions:** Included activities such as building requirements specifications from fictional stakeholders based on case studies and personas. Both GenAI tools and traditional methods were used to create these documents.
- **Analytical Exercises:** Students evaluated and refined both AI-generated and manually created requirements. They used insights from the interactive sessions to improve these documents. They were also asked to exchange requirements documents where some were written manually and others generated by AI, and guess their origins.
- **Reflection Activities:** After each workshop, students were tasked to reflect on their experiences of using GenAI, through both written reflections and group discussions.

All workshops were recorded and transcribed to capture the students' interactions with GenAI.

3.2.2 Surveys

Surveys were distributed to collect additional data about how students felt about using GenAI in the workshops. These surveys were structured with Likert-scale items, which measure levels of agreement or disagreement (Likert, 1932), to evaluate different aspects such as the usability, relevance, and quality of the AI-generated requirements. They were designed to cover different aspects of GenAI's role in RE education, especially the benefits, challenges, and how it improves learning and understanding.

3.3 Data Analysis

The qualitative data were initially analysed separately to create a foundation for the survey design as depicted in Figure 1 below. Later the qualitative results were compared to the quantitative results with the idea of data triangulation.

3.3.1 Qualitative analysis

Thematic analysis was used on the qualitative data from the workshops and reflective sessions. The data was coded iteratively, to identify important themes. It was used to carefully identify, analyse and report patterns in the data (Braun and Clarke, 2006). This method was chosen because it allows for a detailed and nuanced analysis of the data, which helped to understand students' perspectives on using GenAI in RE education.

3.3.2 Quantitative analysis

The data from the surveys were analysed using established statistical methods to show how the responses varied, which highlighted students' evaluations of the benefits and challenges of using GenAI. Descriptive statistics utilized to summarize the data, using measures such as means, medians, and standard deviations (Edling and Hedström, 2003). This approach was selected because it provides a clear summary of the survey results, which makes it easy to interpret and compare responses.

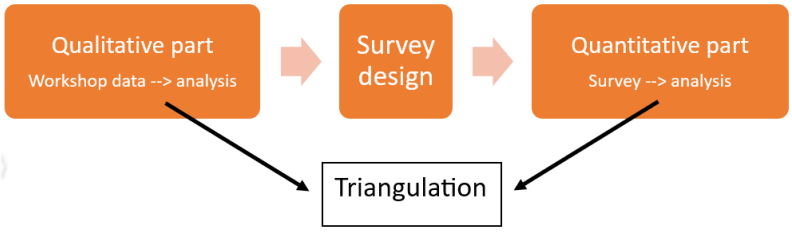


Figure 1: The study's exploratory sequential design

As outlined by Howe (2012), triangulation can be used in mixed method research both to confirm and disconfirm findings, depending on whether data from different data sources converge or diverge. Moreover, this study also used investigator triangulation when findings and claims were discussed in online meetings.

3.4 Ethical Consideration

Informed consent was obtained from each participant included in the study. They were given detailed information about the purpose of the study, their role in the study, and the rights to withdraw at any time (Bryman, 2016).

Confidentiality and anonymity were guaranteed to all participants and maintained throughout the research. Personal identifiers were removed, and pseudonyms were used when sharing the results. In order to align with the guidelines by the Swedish Research Council (Vetenskapsrådet, 2024), all data was securely stored with access that was restricted to the research team.

4. Findings

Findings from the workshop, and findings from the survey are firstly presented separately, and then compared and discussed together.

4.1 Workshop Findings

Twenty-five individuals attended the first workshop, and twelve individuals attended the second workshop. Thematic analysis according to Braun and Clarke resulted in the following three main themes across the workshops: *Facilitating learning, creativity and engagement, Challenges in crafting effective prompts* and *managing AI outputs, and Balancing AI and human insight.*

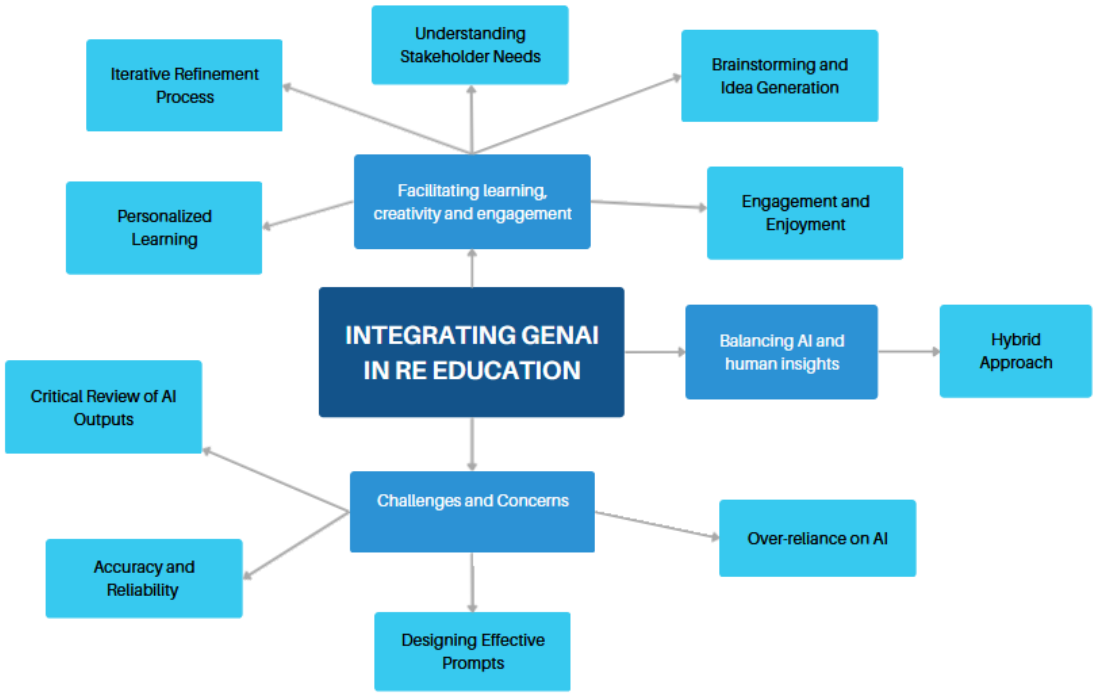


Figure 2: Found themes and their subthemes in the thematic analysis

4.1.1 Facilitating learning, creativity and engagement

In the initial session of the workshops, students were tasked with using ChatGPT to develop requirements based on fictitious case studies and personas. While some students perceived certain AI-generated requirements to be too generic or irrelevant to the context of their case studies, most students appreciated the AI's ability to suggest potential needs and requirements they had not considered on their own. "It was easy to input detailed personas and ask ChatGPT to start from the personas' needs to improve requirements from a user perspective, which made it easier to connect certain traits of personas to relevant requirements and needs," as stated by one student.

Students felt that it was relatively easy to receive useful tips and suggestions on how to improve their requirements documents. Another clear advantage was that ChatGPT could be useful for brainstorming and generating ideas, particularly when something felt unclear or when it was difficult to get started. Several students appreciated AI's ability to suggest features and requirements that they had not thought of themselves, inspiring them to try out new ways of formulating and structuring requirements.

There were several comments that using GenAI was more fun and engaging than manually creating requirements specifications from scratch. The iterative interaction with ChatGPT was especially appreciated by many students. They compared this to iterating requirements in the real industry, where communication with stakeholders often occurs continuously throughout the project. This increased the students' interest and engagement, which was evident in observations of increased levels of active participation. One student noted "When I asked ChatGpt to make the requirements more detailed, they became clearer. I could also see further improvement when I asked for this again, so it helped to point out the same thing several times to get further refinements."

In an exercise where students were tasked with comparing AI-generated requirements with those written by hand, each group could relatively quickly identify which requirements were generated by the AI and distinguish them from those created by other students. It was observed that AI often used more detailed descriptions and more complicated language compared to requirements written by students. However, even if requirements written by students were simpler, they were also easier to understand and had more variations in the language. This process of comparing manually crafted requirements with AI-generated ones helped students understand the strengths and weaknesses of both approaches.

4.1.2 Challenges in crafting effective prompts and managing AI outputs

Despite students' optimism and the potential, they saw with using ChatGPT in their requirements engineering tasks, there were also several challenges. One notable challenge was learning to design effective prompts. A group of students commented that they often felt that generated requirements should have been made up of several smaller parts to make the requirements specification clearer and simpler. However, they acknowledged that this could have been avoided with more specific prompts. One student commented: "You have to spend a lot of time creating your prompts to get useful answers. The output can easily become too general, or on the contrary too specific but in the wrong parts." While some students appreciated that AI could suggest requirements and needs they had not thought of themselves, others viewed this as a disadvantage. Some students noted that ChatGPT could occasionally find some additional aspects that weren't needed or asked for, which forced them to specify their prompts further.

Students noted the importance of critically reviewing the AI-generated output to avoid misleading or incorrect information. They also commented that there could be a risk of becoming too comfortable and relying too much on the AI tool without critically examining the AI-generated information, an aspect that several students identified with. Students also reflected that learning can be negatively affected by directly copying the result and underscored the importance of critically reviewing the output. Students noted that while AI can save time because it is fast and efficient, it is important to still monitor and process everything manually to make sure everything is included and that is ethical and accurate. As one student reflected, "Although we present information, such as personas, to the AI, there is no guarantee that it will use this information correctly to design the requirements. It's important to oversee how the information is used and communicate how it is desired to be used, to ensure that all information is used in an appropriate way."

4.1.3 Balancing AI and human insight

Students argued that despite ChatGPT's extensive knowledge base, humans can still create better requirements specifications. It was pointed out that humans have a better understanding of the values and needs of other humans, which is crucial in software engineering. "Emotions and empathy give us an advantage when it comes to developing systems for other people. AI can't replace humans and should be used as a tool", as noted by one student. Students emphasised that a combination of their own thoughts and AI produced the best requirements specifications. This mixed approach was found to be the most effective strategy because it combines the speed and efficiency of AI with the accuracy of human input.

4.2 Survey Results

The survey was sent to all 35 students who participated in the workshops. Eighteen students responded to the questionnaire with a 51.43% response rate. The answers were scored on a 5-point Likert scale, and the most significant results are listed in *Table 1.* Since all respondents were first-year students, very few had previous experience in requirements engineering. Almost all (94.4%) reported using GenAI tools in their studies or private lives.

Table 1: Survey results

Question	Agree	Part. Agree	Neutral	Disagree	Comp. Disagree	Median	Mean	Std Dev
Higher quality requirements	11.1	55.6	11.1	16.7	5.6	22.3	33.37	29.41
Improved stakeholder needs	5.6	61.1	16.7	16.7	0.0	16.7	33.37	28.7
Provided useful suggestions	27.8	50	5.6	16.7	0.0	16.7	33.37	38.88
Best results with hybrid approach	50	33.3	5.6	11.1	0.0	11.1	33.33	43.36
Improved understanding of RE process	11.1	50	16.7	22.2	0.0	22.2	33.33	24.20
Positive impact on learning	11.1	61.1	5.6	22.2	0.0	22.2	33.33	34.67
Increased engagement & motivation	33.3	38.9	11.1	16.7	0.0	16.7	33.33	33.78
Effort to refine AI-outputs	5.6	44.4	5.6	38.9	5.6	44.5	33.37	24.20
Creating effective prompts	38.9	33.3	0.0	27.8	0.0	27.8	33.33	36.42
Risk of over-reliance	38.9	27.8	11.1	22.2	0.0	22.2	33.33	29.42

Students were asked about their experiences of using ChatGPT to help them create high-quality requirements. A significant majority (83.3%) of the respondents felt that the best results came from combining AI-generated requirements with manually created requirements. Only two (11.1%) disagreed with this statement, while one student had no opinion. More than three-fourths (77.8%) of respondents reported that AI gave useful suggestions that improved their requirements documents, while 16.7% did not feel that the suggestions were useful. Additionally, about two-thirds (66.7%) of respondents reported that ChatGPT improved their ability to capture stakeholder needs and convert them into requirements. However, 16.7% experienced no improvement at all, and an additional 16.7% had no opinion. Regarding using ChatGPT to craft requirements compared to manual methods, 66.7% of respondents reported that writing requirements with AI produced higher quality requirements compared to manual methods. In contrast, 22.3% of respondents felt that crafting requirements by hand yielded the best results, and 11.1% did not know. While most of the respondents reported having a positive experience using ChatGPT to capture and improve requirements, the survey results show some differences in individual experiences with the AI tool.

Students were also asked about their understanding and engagement. A majority (72.2%) of the respondents felt that using GenAI increased their interest and engagement in requirement engineering tasks, while 16.7% did not share this view. In terms of learning and understanding, slightly more than half of the students reported that the use of AI improved their understanding of the requirements engineering process. 22.2% disagreed and 16.7% did not know. Furthermore, 72.2% felt that ChatGPT contributed positively to their overall learning experience, while 22.2% had a different opinion.

Challenges were explored by asking the students about writing prompts and managing AI outputs. Of the respondents, 72.2% assessed it to be challenging to create effective prompts that produced the desired result, of which 38.9% expressed that it was very difficult. The others experienced no difficulty in creating effective prompts. Regarding AI-generated output, about half of respondents reported that it was often time-consuming and required effort to refine and adjust suggestions generated by the AI. One student had no opinion, and the rest did not find it to be frustrating to rework AI-generated answers. Researchers also asked about students' experiences of over-reliance on AI-generated output, which can negatively impact the quality of the

requirements and the learning process. Two-thirds (66.7%) of respondents acknowledged that this was a problem, with 22.2% disagreeing.

5. Discussion

The aim of this study was to explore the impact of GenAI on the pedagogical aspects of higher education RE, and how it might influence students' learning experiences. This was investigated by developing a series of workshops within the course Analysis and Design of Information Systems, followed by a questionnaire. Our results show that while GenAI can enhance students' abilities to understand and translate stakeholder needs into good requirements by providing valuable suggestions and inspiration, many students reported challenges in writing effective prompts and refining AI-generated answers to meet specific needs.

Overall, students found that GenAI gave useful suggestions and inspiration through iterative interactions that mimicked realistic communication with stakeholders, suggesting that GenAI has the potential to make higher education more engaging and interactive. This aligns with previous research (Michel Villarreal et al., 2023) about GenAI's contribution to enhance education by making it more engaging and immersive because of its dynamic and interactive characteristics. In addition to assisting with managing requirements, students found AI useful for brainstorming to help them generate ideas or get started. This was evident from both the workshop and the survey results, and corroborates findings by Moravánszky (2024) that AI can be helpful as a brainstorming partner to grapple with challenges in refining requirements. Additionally, the survey results show that GenAI improved students understanding for the requirements engineering process and contributed to their overall learning experience. This supports previous research (Hartley, 2023; Daun and Brings, 2023) that indicate that personalized learning through AI can improve student engagement and motivation. This demonstrates possibilities in how AI-tools can be implemented and adapted in different educational contexts.

Despite the positive aspects, there were also several challenges identified. It was notable that students faced difficulties in creating effective prompts, which is a critical factor for success when using GenAI in RE (Ronanki et al., 2024). There were several students who felt that it was time-consuming with repeated interactions where they had to guide the AI to refine and adapt requirements until they fulfilled specific needs. Another challenge was the risk of becoming too dependent on AI, which could lead to diminishing students own critical thinking and problem-solving skills (Kasneci et al., 2023). This indicates a need for training and supporting students in creating effective prompts and reviewing AI-generated content, which aligns with previous research (Petrovska et al., 2024, Moravánszky, 2024) that emphasizes the importance of guiding students in how to use AI-tools, especially in designing prompts and developing their abilities to critically reviewing AI-generated content.

Students favoured a combination of using both AI-generated and manually created requirements. This mixed approach takes advantage of the speed and efficiency of the AI as well as human insight and understanding to ensure that the requirements are relevant and tailored after the stakeholders' needs. This balance is essential in order to maximize the advantages of AI without losing the valuable insights that human traits such as empathy and critical thinking can provide, as also suggested by Shaji George (2023). While students recognized AI as valuable support, they emphasised that the tool should be used to complement their work rather than replace it entirely.

6. Conclusion and Future Work

To answer the research question, our findings show that students generally had a positive experience of using GenAI in RE education. They valued its efficiency and found it useful for tasks such as brainstorming and initiating drafts, which made the learning process more engaging and helped with idea generation. However, our study also revealed challenges. Students often faced difficulties in writing effective prompts and felt it was time-consuming to refine AI-generated outputs. Additionally, the risk of over-reliance on AI-generated content was highlighted, which points to the need for critical review to ensure accuracy. A mixed approach, combining the benefits of AI with human oversight, proved to be the most effective strategy. To summarise, GenAI showed a promising potential to add value to requirements engineering in higher education, but it requires a careful implementation to improve the learning outcomes.

This study was carried out in relatively small and heterogeneous group, and an interesting future follow-up would be to repeat the study with a larger and more diverse sample. What authors see as the most interesting and more challenging continuation would be to develop strategies for effective and ethical integration of GenAI in teaching and learning.

References

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3, 77–101. https://doi.org/10.1191/1478088706qp063oa.

Bryman, Alan. (2016). *Social Research Methods* (5th ed.). Oxford University Press.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43. https://link.springer.com/article/10.1186/s41239-023-00411-8

Choudhuri, R., Liu, D., Steinmacher, I., Gerosa, M., & Sarma, A. (2024). How Far Are We? The Triumphs and Trials of Generative AI in Learning Software Engineering. In: I*CSE 2024 Research Track*, 17 April 2024, Luis de Freitas Branco - Generative AI Studies. https://arxiv.org/abs/2312.11719.

Creswell, J.W. & Plano Clark, V.L. (2017). *Designing and Conducting Mixed Methods Research*. 3rd ed. Thousand Oaks, CA: SAGE Publications.

Daun, M. & Brings, J. (2023). How ChatGPT Will Change Software Engineering Education. *ITiCSE 2023: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education*, 1, 110–116. https://doi.org/10.1145/3587102.3588815.

Daun, M., Grubb, A.M., Stenkova, V., & Tenbergen, B. (2022). A systematic literature review of requirements engineering education. *Requirements Engineering, 27*, 1–24. https://doi.org/10.1007/s00766-022-00381-9.

Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., ... & Sarsa, S. (2024). Computing education in the era of generative AI. *Communications of the ACM*, *67*(2), 56-67.

Edling, C. & Hedström, P. (2003). *Kvantitativa metoder - Grundläggande analysmetoder för samhälls- och beteendevetare*. Lund: Studentlitteratur AB.

Hartley, S. (2023). More on AI Generated Content. *Current Psychiatry, 22* (6). https://doi.org/10.12788/cp.0371

Howe, K. R. (2012). Mixed methods, triangulation, and causal explanation. *Journal of Mixed Methods Research*, *6*(2), 89-96. https://doi.org/10.1177/1558689812437187

Kadaruddin, K. (2023). Empowering Education through Generative AI: Innovative Instructional Strategies for Tomorrow's Learners. *International Journal of Business, Law, and Education*, 4(2), 618 - 625. https://doi.org/10.56442/ijble.v4i2.215

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. L*earning and Individual Differences*, 103, 102274. https://www.mdpi.com/2227-7102/13/4/410

Likert, R.A., 1932. Technique for the measurement of attitudes. *Archives of Psychology*, 140: 1-55.

Memon, R.N., Ahmad, R., & Salim, S.S. (2010). Problems in Requirements Engineering Education: A Survey. *FIT '10: Proceedings of the 8th International Conference on Frontiers of Information Technology*, Article No. 5, 1-6.

Moravánszky, A. (n.d.). *Doctoral School of Education, University of Szeged, Hungary & University of Applied Sciences of the Grisons, Switzerland*. https://ceur-ws.org/Vol-3672/EuT-paper2.pdf.

Nolen, A.L. & Putten, J.V. (2007). Action research in education: Addressing gaps in ethical principles and practices. *Educational Researcher, 36*(7), 401-407.

Norton, L. (2009). *Action Research in Teaching and Learning: A Practical Guide to Conducting Pedagogical Research in Universities.* Routledge.

Ochodek, M., & Kopczyńska, S. (2018). Perceived importance of agile requirements engineering practices—a survey. *Journal of Systems and Software,* 143, 29-43.

Petrovska, O., Clift, L., Moller, F., & Pearsall, R. (2024). Incorporating Generative AI into Software Development Education. *CEP '24: Proceedings of the 8th Conference on Computing Education Practice*, 37-40. https://doi.org/10.1145/3633053.3633057.

Rajbhoj, A., Somase, A., Kulkarni, P., & Kulkarni, V. (2024). Accelerating Software Development Using Generative AI: ChatGPT Case Study. In: *17th Innovations in Software Engineering Conference (ISEC 2024) Research Papers*, 23 February 2024, Room 1 - R103 - Session 2 - LLMs Meet SE.

Ronanki, K., Cabrero-Daniel, B., Horkoff, J., & Berger, C. (2024). Requirements Engineering Using Generative AI: Prompts and Prompting Patterns. In: Nguyen-Duc, A., Abrahamsson, P., & Khomh, F. (eds.) *Generative AI for Effective Software Development.* Springer, Cham. https://doi.org/10.1007/978-3-031-55642-5_5.

Shaji George, A. (2023). The Potential of Generative AI to Reform Graduate Education. *Partners Universal International Research Journal.* https://doi.org/10.5281/zenodo.10421475

Spichkova M (2019) Industry-oriented project-based learning of software engineering. In: 2019 24th International conference on engineering of complex computer systems (ICECCS), pp 51–60

Sullivan, M.; Kelly, A.; & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning and Teaching*, 6, 1-10. https://ro.ecu.edu.au/ecuworks2022-2026/2501/

Vetenskapsrådet. (2014). Conducting ethical research. https://www.vr.se/english/applying-for-funding/requirements-terms-and-conditions/conducting-ethical-research.html [20 June 2024].

Yeralan, S., & Lee, L. A. (2023). Generative AI: Challenges to higher education. *Sustainable Engineering and Innovation*, *5*(2), 107-116.

Yu, H. (2023). Reflection on whether ChatGPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14, 1181712. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1181712/full

Late Submissions

Challenges in AI Implementation: Perspectives from Practice and Research

Clemens Kerschbaum[1,2] and Raphael Dachs[1]
[1]Argus-AI GmbH, Austria
[2]New Design University, Austria

clemens.kerschbaum@argus-ai.at
raphael.dachs@argus-ai.at

Abstract: Artificial Intelligence (AI) has become an inevitable topic for organizations across various sectors and sizes, offering promising applications as technological accessibility continues to expand. Despite its potential, practical implementation of AI-based Systems remains difficult with particular challenges tied to specific organizational contexts. Often companies invest heavily in AI development but encounter problems such as failure to achieve market readiness of the prototype or the systems struggling to deliver expected benefits. These setbacks often stem from flawed implementation strategies, excessive reliance on technology, or inadequate integration into existing organizational frameworks. Therefore, this paper addresses these challenges encountered at different phases of AI implementation projects. To this end, we initially conduct a Rapid Structured Literature Review (Armitage and Keeble-Allen, 2008), examining the literature on AI implementation cases and associated scholarly reviews. Extending the initial analysis, we experiment with AI driven document analysis as a means to integrate findings of a greater amount of publications into the review. The literature review is subsequently complemented by insights from our own consultancy experiences from the field of AI consultancy. The paper gives an overview of the most salient challenges in AI implementation projects and points out some approaches to mitigate those challenges. From a methodological standpoint it shows that AI driven reviews can yield similar results as conventional reviews, but may lack some explanatory depth. We find that a combination of manual and automated approaches tends to be the most effective strategy.

Keywords: AI implementation, Machine learning, Organizational challenges, Semi-automated literature review, AI security

1. Introduction

According to a recent publication by Gartner, "*Investment in AI has reached a new high with a focus on generative AI*"[1]. Yet at the same time, in Gartner's infamous Hype Cycle, the topic of Generative AI is at the peak of inflated expectations, likely nearing the edge of disillusionment as more and more organizations encounter issues when they try to implement AI into their undertakings . Although the technology has generated dazzling success stories for some early adopters, its suitability for the masses appears more and more questionable, as reports of failed AI initiatives are increasing. Many authors claim that AI technology seems to fall short of delivering the anticipated business value it promised, or fails to be operationalized, thus often remaining in a prototypical stage (e.g. Hopf et al., 2023). Thus, elucidating the issues that organizations encounter when trying to adopt AI technology seems an interesting avenue for research. And indeed the number of publications on AI implementation challenges is growing rapidly in many areas. For practitioners, however, it can be challenging to transfer those findings to their own endeavors, as scientific literature is not always easily accessible and the interpretation or generalization of often very specific studies poses another hurdle. Thus, with this review, we attempt to take a step of abstraction and analyze typical challenges in order to build categories of common pitfalls in AI implementation projects. This is beneficial for scientific knowledge transfer into practice but can also yield useful results for academia, as it reveals potential open ends or blind spots of the current discourse. Not least, recognizing common categories of issues paves the way to the development of appropriate solutions that will also be briefly addressed in the paper. Our analysis adapts a combination of manual and automated methods to identify AI implementation challenges from the literature. Therefore, In addition to addressing the content dimension, this article also enhances our understanding of AI's potential as a research tool for analyzing scientific literature and extracting relevant information.

2. Method

In order to examine the issues that organizations encounter when dealing with AI and its implementation into their operations, we conduct an exploratory literature review, adapting Armitage and Keeble-Allen's (2008) Rapid Structured Literature Review (RSLR) framework. Complementing this analysis, we also run an automated analysis over a greater sample of literature to cross check our findings against other publications. Where

[1] https://www.gartner.com/en/documents/5505695

possible, we also put the results in the context of our practical experience from our consulting work in the AI domain. Thus the research presented herein consisted of three steps: the first is an analysis of the literature and the extraction of common categories of issues that organizations would encounter according to the surveyed publications. The second step consists of a screening of a large sample of publications on AI implementation and respective challenges and the automated generation of categories of challenges to cross-check our initial findings. Lastly, we screen our own process documentations from about ten years of consultation projects. We then compare and contrast the data with the aim to integrate the findings from all sources.

For the manual review of the literature, we opted for Armitage & Keeble-Allen's methodological framework for conducting a RSLR (Armitage and Keeble-Allen, 2008; Armitage and Keeble-Ramsay, 2009), that is in alignment with our goal to get an overview of the rapidly growing literature on AI implementation issues. A Rapid Structured Literature Review (RSLR) consists of six stages: defining a research question, developing a search strategy (databases and keywords), conducting the literature search, screening and selecting relevant studies, extracting and synthesizing data, and reporting the findings in a structured format.

2.1 Research Question

In this phase, the most important aspect to consider is which facets of the research question are of particular importance to achieve a desired research output. For our purposes, we knew that the focus should be on AI projects, meaning that we wanted to address initiatives where an organization would try to implement AI into their undertakings, but exclude all organizations who rely on AI as their core business. What is more, our goal was to adopt a broad perspective on the challenges of AI implementation in order to uncover the true heart of the matter. This idea stems from practice, where we often found that its hardly purely technical issues that hindered project success but a diffuse combination of various aspects.

With the formulation "development and implementation of AI systems", we scope our search towards customized AI applications that go beyond the use of easily accessible AI tools such as Chat GPT and alike. Ultimately, these considerations led to the formulation of the following research question as a guideline for our study:

What are key problems and challenges, encountered in the development and implementation of AI systems within companies and other organizations, that hinder AI project success?

2.2 Identifying Relevant Studies

According to our goal to include a great variety of sources and ideally a mixture of case reports and, if available, existing reviews of AI implementation challenges, we initially aimed for multiple databases. The focus, however, was set to papers published in academic outlets to ensure a certain degree of objectivity and reliability of the data. We thus decided to use the databases of Scopus, ProQuest and Google Scholar to search for relevant publications using the search strings "AI implementation" AND challenges AND project (AND "literature review"); "machine learning implementation" AND challenges (AND "literature review").

Not surprisingly, the initial search yielded five-digit numbers of results, which confirmed the assumption that there is already a large number of publications and might indicate the relevance of the topic. However, numbers alone are no reliable indicator for relevance, as the whole field of information systems is extensive and relevant papers easily exceed tens of thousands (Wagner et al., 2022).

As a first means to filter out papers for the subsequent analysis, we went through the titles of those papers that the search engines marked as the most relevant and collected them in a separate folder to prepare the next step of the analysis. In addition to the systematic keyword search, we also added papers to our database that we either encountered before this research project or that we found by browsing the references of other papers using the snowball method. A noteworthy observation during the literature search was that the results from the online search engines produced some overlap, but each provided mostly unique results.

2.3 Study Selection

As the initial search led to a very large number of papers to be potentially included in the review, we were thinking of either reducing the sample through inclusion / exclusion criteria as suggested by the RSLR methodology (Armitage and Keeble-Allen, 2008; Armitage and Keeble-Ramsay, 2009), or automating the analysis using a language model to search for content that would help to answer the research question. Ultimately, we opted for a mixed approach: we filtered our sample for all those publications that best matched

the original query for manual reading to get an idea of reoccurring topics and create a framework for the analysis of the remaining body of literature. The remaining sources were analyzed automatically using the platform Scispace[2], an application that allows for searching and analyzing scientific literature using large language models. For the initial manual analysis and development of the categories that guided the following analysis we used the following sources: Brock and Von Wangenheim (2019), Gabsi (2024), Lee et al. (2023), Mergel et al. (2023); Oldemeyer et al. (2024), Reim et al. (2020) and Viskova-Robertson (2023). The overall sample for the automated analysis consisted of 97 articles.

3. Automated Analysis

After the manual analysis of the aforementioned publications and the development of the categories as depicted in Table 1 and Table 2 in the columns titled "initial Dimensions", we began the automated analysis by uploading our sample of publications to the research platform Scispace (available via www.typeset.io). The application offers various features, including the automated analysis of pdf documents which we utilized for our sample. We analyzed the literature with regards to challenges and respective guidelines to overcome those. In order to extract the relevant data we ran a query that would summarize the key findings regarding AI implementation challenges of all the papers in our library. As a next step, we built thematic clusters on the basis of the summarized findings. This led to ten dimensions of AI implementation challenges as depicted on the right hand side in Table 1 and Table 2. After the analysis for implementation challenges led to promising results, we conducted the same procedure for proposed solutions in order to compare if challenges and solutions would match, as an indicator for reliability of the analysis.

4. Findings (Challenges and Potential Solutions)

In this section we present and overview of challenges and proposed solutions that we identified from the surveyed literature. We compare the dimensions of the initial manual analysis with the analysis using AI technology.

4.1 Challenges of AI Implementation

We first define six dimensions (Knowledge & Expertise, Financial Considerations, Data, Organizational Context, Technical issues, Security & Reliability) describing the types of challenges, organizations encounter when implementing AI that we derive from the manually surveyed articles. Then we progress to compare our own analysis with the automated version based on extracted sentences and keywords that describe AI implementation challenges in the remainder of the literature. Table 1 compares our initial dimensions, with the results of the automated analysis.

Table 1: Comparison of initial dimensions and automatically generated Dimensions

Initial Dimensions	Automated Analysis
Data	Data-related Issues
Security & Reliability	Ethical and Societal Concerns
	Transparency and Explainability Concerns
Technical Issues	Technological Barriers
	Adoption and Implementation Challenges
Organizational Context	Management and Governance Issues
	Integration with Existing Processes
	Cultural and Leadership Challenges
Knowledge & Expertise	Workforce and Skills Gap
Financial Considerations	Resource Limitations

The table shows that the overarching dimensions map fairly well, although the automated analysis yielded more dimensions than the manual analysis. When going into detail, we found that the overlap of identified challenges was even more compelling. Table 2 presents our analysis with all subcategories included and

[2] www.typeset.io

compares it to the results from the automated analysis, also including all subcategories. The automated analysis yielded very few results that were obviously out of place (e.g. the statement "Exclusion of relevant papers due to specific keywords used") which we removed to keep the table concise. We also removed a few results that were too domain specific, resulting from the sample that contained many articles from specific domains (e.g. "challenges in developing construction robots due to the constantly changing environment"). Apart from these minor corrections, we depict the results as is. We found that the automatically generated subcategories are not as selective as the manually generated ones. Some subcategories in the automated column can be found twice in different dimensions. Yet, arguably the results from the automated analysis are pretty satisfactory when compared to the manual analysis, when the goal is to provide an overview of the challenges encountered.

Table 2: Detailed comparison of manual and automated analysis, including subcategories

Initial Dimensions	Initial Subcategories	Automated Dimensions	Automated Subcategories
Data	Data quality Availability of data Fragmented / Limited Training Data Data security Data Privacy Limitations through analog processes	Data-related Issues	Data quality Data privacy and security concerns Data management Lack of dataset availability Limited access to data Data ethics
Security & Reliability	related to the Data Dimension; two subcategories: A) Technical reliability & security Privacy concerns in public administration contexts Sensitive data related to processes, products and customers Particular challenges of regulated and/or critical industries (adherence to standards) e.g. aviation, medicine, infrastructure,… Hardware requirements Biases in algorithms B) Ethical considerations Transparency and traceability Fairness & equity (especially in public admin.) Potential influence on societal values Accountability and responsibility	Ethical and Societal Concerns	Ethical considerations Ethical issues Algorithmic bias Privacy violations Discrimination Trust in AI technology Transparency and accountability Societal inequalities Fair access to transformative technology
		Transparency and Explainability Concerns	Lack of explainability in AI models Algorithm opacity Insufficient documentation and traceable logs
Technical Issues	Maintenance & ensuring continuous performance Scalability & Flexibility to accommodate changing requirements Interoperability, Integration of digital & existing technology AI implementation as a trial-and-error process Availability of technology (no off the shelf solutions) requires customized	Technological Barriers	Technical feasibility, regulatory norms, privacy concerns, and ethical concerns Integration challenges Technological challenges in developing construction robots due to the constantly changing environment High cost of implementation Non-standardization hinders automation and robotics efficiency

	solutions	Adoption and Implementation Challenges	Implementation barriers at tactical, operational, and strategic levels Lack of strategy to implement AI Resistance from employees Ambiguity in decision-making with AI
Organizational Context	Management support / resistance from management Resistance / support from workforce Trust & Accountability for failure Lack of leadership towards digitalization Internal / structural resistance to change Lack of organizational agility; emergent nature of AI projects (Brock und Von Wangenheim, 2019, p. 128) Challenging for SMEs (Oldemeyer et al., 2024, p. 24) Mutual misunderstandings of AI capacities in different departments / roles Productivity evaluation Complexity & effects on the organization Restructuring of organizational activities (transformational aspect)	Management and Governance Issues	Governance, scalability, and privacy issues related to AI implementation Lack of organizational capabilities related to data Lack of defined requirements for generating and maintaining records Organizational inertia
		Integration with Existing Processes	Alignment with existing processes and systems Compatibility issues with traditional business models
		Cultural and Leadership Challenges	Organizational culture Leadership challenges Misconceptions related to human and machine compatibility
Knowledge & Expertise	Availability of skilled workforce Upskilling of surrounding functions General understanding of AI in the organization Lack of example cases, hardly any empirical advice Little experience of consequences of AI Misconceptions of AI possibilities & limitations	Workforce and Skills Gap	Addressing workforce skills gap Lack of skilled personnel Insufficient employee qualification Human cognitive flexibility and capacity for learning skills Employee training and upskilling Knowledge and skills gap
Financial Considerations	High initial investments Sunk costs for abandoned projects Difficulties if expected results do not (yet) materialize	Resource Limitations	Budget constraints Resource constraints Limited access to AI expertise Infrastructure limitations

Looking at the findings themselves, we find that although it is possible to create distinct categories, those are strongly interdependent. An isolated view of one category or a respective intervention thus poses the threat of unintended consequences in one or more other dimensions. This illustrates the complexity that is mentioned as part of the organizational context dimension, which is mentioned in multiple publications. AI implementation might require organizations to revise their organizational structure charts (Viskova-Robertson, 2023), and on the other hand, in contexts like the public sector, where organizations cannot simply adapt their structure, AI initiatives need to be aligned with the organizational context (Mergel et al., 2023). Literature confirmed the relatively subordinate role that technical challenges play when compared to other aspects such as Knowledge and expertise. However, the related aspect of Data availability and quality is of utmost importance. However, regarding Data, we once more found strong ties to other dimensions, particularly the organizational context as well as Security & reliability and Ethical & Privacy considerations.

4.2 Proposed Solutions

After the analysis of challenges, we were curios to see what solutions the sample had to offer in comparison to the challenges. We analyzed the full sample of papers and extracted short sentences and keywords from the data. Those were subsequently clustered taking into consideration their frequency of occurrence and explanatory power.

The analysis showed that the most prominent suggestion concerned the issue of data management and quality which includes organizational capabilities to collect data in a systematic way and implement standardized documentation practices for data and respective models. The second cluster, matching the findings from the analysis of challenges concerns the acquisition of knowledge and skills development. Here, literature suggests continuous training and upskilling of employees in all aspects of AI related issues (technical, ethical, legal, managerial aspects) including the requirement for AI literacy training by educating bodies. The next important area to consider is to adapt a mid-, to long term perspective on AI. This means to develop a comprehensive AI strategy that covers both required technical steps as well as a strategy to develop organizational capabilities for AI implementation. Further, an AI strategy should deal with the alignment of AI with stakeholders needs and values. Recommendations for such comprehensive AI strategies indicate the interconnectedness of the challenges from an organizational perspective. Having a well done strategy in place that includes appropriate measures for project success at a given time can mitigate the aforementioned risk of abandoning AI initiatives too early, when expected results do not materialize at first. Closely tied to that requirement is inter-organizational communication and collaboration, especially between IT and business departments. Literature calls for collaborative problem-solving, trust and establishing clear communication channels. Trust is also a reoccurring topic with regards to the systems itself, meaning that whatever the use case, respective AI applications should be as transparent as possible and results should be reproduceable. Additionally, stakeholders should be involved in the process of implementation where possible. A separate category, however, closely tied to trust is ethics and governance, where literature recommends ethics- and privacy-based audits along with mechanisms to ensure transparency, accountability, and contextually relevant AI deployment. Considering the frequency of mentions, technical considerations (excluding data-related aspects) only appear at the bottom of the list.

Reading through this suggestions again illustrates the complexity of AI initiatives and the need for easily applicable frameworks or guidelines for practitioners. Some authors have already worked out respective approaches. Viskova-Robertson (2023) or Akbarighatar et al. (2023), for example suggests to understand AI systems as socio-technical systems. Lee et al. (2023) propose an Input-Process-Output model as the research framework to investigate the antecedences, processes, and consequences of AI implementation in organizations. Other examples include the TOE framework (Shahzadi et al., 2024) for AI in supply chain management, DIGITAL guidelines for successful AI Applications (Brock and Von Wangenheim, 2019) or a roadmap for Business Model Innovation in the context of AI (Reim et al., 2020). In the field of public administration, innovation labs or specified AI labs providing safe environments for emerging AI technologies have also proven successful (Mergel et al., 2023).

4.3 Limitations

Due to the somewhat experimental research design, we acknowledge that there are several limitations to the results of this study. First, although we searched only for scientific literature we did not perform any deeper quality assessment of the selected articles, however, we excluded articles that appeared as mere essays or opinion pieces or did obviously not meet academic standards. Second, as we derived initial categories of challenges from a relatively small sample of papers, we cannot assure with certainty that these categories would have emerged the same, had we started from different base-articles. Third, the way we conducted the analysis definitely has the advantage that it can take into account a very large body of literature, yet we must assume that a great share of context is lost through such automation. What is more, it is difficult to maintain traceability of individual arguments throughout the analysis as the approach tends to aggregate data into higher-level statements. While this can be acceptable for the purposes of a meta level exploratory study that aims to provide an overview of a large body of literature, it is definitely an issue to be addressed for more systematic and structured types of methodologies. From a researchers individual perspective, another interesting observation was that designing and experimenting with prompts for the analysis brutally uncovers the lack of “intentionality” of a large language model (i.e. the model does not share researchers curiosity and interest for the discovery of insights). From an objective standpoint, this is of course unsurprising, but it has the implication for research practice that intentionality and the specific research interest need to be

incorporated into the prompts to achieve satisfactory results. Table 3 provides an example for the development of the query regarding possible solutions to overcome AI implementation challenges.

Table 3: Developing a prompt to analyze solutions to AI challenges in the sample using typeset.io

Solution Keywords v1	Solution Keywords v2	Solution Keywords v3	Solution Keywords v4
Provide keywords for solutions that the paper suggests to overcome the challenges presented in the paper.	Provide keywords for solutions that the paper suggests to overcome the challenges presented in the paper. Exclude everything that is domain specific. Exclude everything that has to do with medicine.	What are key problems and challenges, encountered in the development and implementation of AI systems within companies and other organizations, that hinder AI project success? Provide keywords for what the paper suggests to overcome these challenges presented in the paper. Exclude everything that is domain specific. Exclude everything that has to do with medicine.	What are key problems and challenges, encountered in the development and implementation of AI systems within companies and other organizations, that hinder AI project success? Give me keywords for what the paper suggests to overcome these challenges presented in the paper. Exclude everything that is domain specific. Exclude everything that has to do with medicine. Include only the proposed solutions, disregard keywords that only describe challenges

5. Conclusion

From our analysis of the literature and experience in consulting work, we can conclude that the challenges of AI implementation go beyond technical aspects and are in fact very diverse. However, the most prominent challenge for the successful implementation of AI, appears to be knowledge and expertise. This includes technical expertise but more importantly, a general understanding of the technology and its functioning beyond the specialized experts in IT departments. Understanding AI in that sense serves as a gatekeeper to the meaningful consideration of other issues which all rely on the organizational context that ultimately shapes the landscape for AI implementation. Here, the most salient challenges are financial constraints, availability and quality of data as well as security and reliability of the systems (especially in critical contexts). Challenges embedded in the organizational context also include support of management and leadership towards AI, sufficient agility of the organization and the complex issue of process- and eventually business model transformation elicited through AI implementation. Yet, it is noteworthy that not all AI initiatives must necessarily re-invent the wheel as “realistic AI” is said to prevail for quite some time, before technology will take over everything (Brock und Von Wangenheim, 2019, p. 129). When looking at proposed solutions to the challenges identified, we see a trend towards holistic frameworks (e.g. Brock and Von Wangenheim, 2019), taking a socio-technical systems perspective (Viskova-Robertson, 2023) that acknowledge the complexity of AI implementation and interconnectedness with other organizational domains. A general recommendation derived from these findings could be for organizations to prioritize foundational training in AI for all employees, with the aim of establishing a shared understanding and common knowledge base. This could benefit the systemic integration of AI within organizations, making expectations of the technology more realistic while also enabling a better recognition of its potential.

In further research, it would be interesting to gather and analyze empirical evidence of AI projects that take up such an integrated perspective on AI and see how such projects turn out in comparison to previous examples. Additionally, research could address the issue of AI literacy from a general perspective and develop strategies to convey an essential understanding of the technology to a broad audience.

From a methodological perspective, it becomes evident that large language models can indeed produce useful results for scientific reviews. However, at least in our case, the findings tended to be more abstract compared to the manual analysis. It also became clear that researchers must formulate their queries very precisely in order to obtain results that align well with the specific research objectives. Despite some minor inconsistencies that may have eventually been caused by less-than-optimally phrased inputs, we argue that the method of automated literature analysis can be useful for exploratory literature studies. Yet, from our experience from this trial, it might be less suited for in depth analyses where a more nuanced understanding of the content is required.

Note: More examples and data from the analysis as well as the full list of references underlying the analysis are available from the authors. The full list of references is also available here < https://shorturl.at/0IKRu > . Since this type of review is a novel approach, we highly welcome feedback and suggestions for improvements of the methodology.

References

Akbarighatar, P., Pappas, I., Vassilakopoulou, P., 2023. A sociotechnical perspective for responsible AI maturity models: Findings from a mixed-method literature review. Int. J. Inf. Manag. Data Insights 3, 100193. https://doi.org/10.1016/j.jjimei.2023.100193

Armitage, A., Keeble-Allen, D., 2008. Undertaking a structured literature review or structuring a literature review: Tales from the field. Electron. J. Bus. Res. Methods 6, 103–114.

Armitage, A., Keeble-Ramsay, D., 2009. The rapid structured literature review as a research strategy. US-China Educ. Rev. 6.

Brock, J.K.-U., Von Wangenheim, F., 2019. Demystifying AI: What Digital Transformation Leaders Can Teach You about Realistic Artificial Intelligence. Calif. Manage. Rev. 61, 110–134. https://doi.org/10.1177/1536504219865226

Gabsi, A.E.H., 2024. Integrating artificial intelligence in industry 4.0: insights, challenges, and future prospects–a literature review. Ann. Oper. Res. https://doi.org/10.1007/s10479-024-06012-6

Hopf, K., Müller, O., Shollo, A., Thiess, T., 2023. Organizational Implementation of AI: C RAFT AND M ECHANICAL W ORK. Calif. Manage. Rev. 66, 23–47. https://doi.org/10.1177/00081256231197445

Lee, M.C.M., Scheepers, H., Lui, A.K.H., Ngai, E.W.T., 2023. The implementation of artificial intelligence in organizations: A systematic literature review. Inf. Manage. 60, 103816. https://doi.org/10.1016/j.im.2023.103816

Mergel, I., Dickinson, H., Stenvall, J., Gasco, M., 2023. Implementing AI in the public sector. Public Manag. Rev. 1–14. https://doi.org/10.1080/14719037.2023.2231950

Oldemeyer, L., Jede, A., Teuteberg, F., 2024. Investigation of artificial intelligence in SMEs: a systematic review of the state of the art and the main implementation challenges. Manag. Rev. Q. https://doi.org/10.1007/s11301-024-00405-4

Reim, W., Åström, J., Eriksson, O., 2020. Implementation of Artificial Intelligence (AI): A Roadmap for Business Model Innovation. AI 1, 180–191. https://doi.org/10.3390/ai1020011

Shahzadi, G., Jia, F., Chen, L., John, A., 2024. AI adoption in supply chain management: a systematic literature review. J. Manuf. Technol. Manag. https://doi.org/10.1108/JMTM-09-2023-0431

Viskova-Robertson, A., 2023. Socio-Technical System and Organizational Ai Integration: An Integrative Literature Review. Acad. Manag. Annu. Meet. Proc. 2023, 1–6. https://doi.org/10.5465/AMPROC.2023.48bp

Wagner, G., Lukyanenko, R., Paré, G., 2022. Artificial intelligence and the conduct of literature reviews. J. Inf. Technol. 37, 209–226. https://doi.org/10.1177/02683962211048201

Driving the Future: AI in Transportation

Isabella Xu[1], Murtaza Haque[1] and Song Fu[2]
[1]University of North Texas Denton, USA
[2]Department of Computer Science and Engineering, University of North Texas Denton, USA

isabella.g.xu@gmail.com
murtaza0713@gmail.com
song.fu@unt.edu

Abstract: Transportation lies at the heart of our society, shaping nearly every aspect of our lives. Through its most fundamental function—mobility—transportation upholds the interconnected nature of our world. Economically, transportation controls the distribution of products, enabling mass production and global trade. It also expands personal mobility, expanding access to jobs and fostering communal connectivity. Culturally and socially, transportation diffuses ideas, values, and customs while providing access to essential services like medical care and education. Given these profound impacts, it is clear that the transportation industry has undergone drastic changes since its inception, often driven by single innovations that redefine mobility. Consider the invention of the steamboat in 1787. Upon its introduction, passenger travel became widely accessible, enabling a stream of ideas and culture to diffuse across nations. Simultaneously, it transformed commercial shipping by reducing the time and cost required of transporting cargo. Modern means of transportation have evolved to a network of defined highways and roadways. To keep order in this network, our transportation system requires a solution – one capable of handling its diverse array of challenges and complexities. Enter Artificial Intelligence (AI). AI-driven solutions within the transportation system offer a dynamic and flexible approach. With the ability to continuously adapt to the fast-moving pace of our transportation network, AI has garnered the attention of many. In recent years, this growing interest has resulted in the acceleration of research at the intersection between AI and transportation, signifying the beginning of a tremendous shift in the way mobility is perceived. The ever-increasing demand for transit in our society brings forth pressing issues such as environmental decay, traffic congestion, and safety risks. This paper will delve into AI applications that hold enormous potential for addressing these challenges. This paper will also explore the multifaceted implications of AI in transportation, providing a detailed overview of its impacts arising in areas of efficiency, safety, and sustainability. Then, taking a step back, the paper will outline broader implications for our society, specifically how AI-enabled changes in the transportation industry will impact society at large. Key considerations surrounding implementation in these sectors will be presented, outlining effective strategies associated with AI-enabled transportation. The adoption of new technology in our society is normally accompanied by various risks. Unfortunately, AI in transportation is no different. To ensure that the arrival of AI-enabled transportation does not take a turn for the worse, this paper will consider potential repercussions in data security, algorithmic bias, and user privacy. Furthermore, it will discuss the adequacy of current regulations related to AI transportation. Lastly, the paper will focus on the future direction of AI in furthering the reliability of transportation systems. It will present information regarding research gaps and future goals, all with the aim of demonstrating the multitude of opportunities unfolding in the transportation industry. In doing so, this paper intends to demystify the seemingly “intimidating” sector of AI in transportation while emphasizing the importance of accountability in its integration.

Keywords: AI in transportation, Autonomous systems, Traffic management, Ethical implications, Optimization

1. AI Applications in Transportation

The field of AI reveals endless opportunities for innovation in transportation. With its ability to analyze vast amounts of data, adapt to evolving conditions, and uncover hidden patterns in traffic flow, AI stands to make an immense impact on the transportation network. Its applications span a wide range of transportation infrastructure elements, from traffic flow pattern analysis to the control and management of personal vehicles. As AI technology continues to advance, it's set to change the very landscape of our transportation networks. Let’s delve into the multitude of ways in which AI will shape our transportation ecosystem.

1.1 Route Optimization

Navigating through the complex network of transportation is often a time–consuming task. Consider all the different factors influencing the flow of traffic: route closures, congestion, weather conditions. Even a slight change in any of these elements can cause fluctuations in transit times. In this way, the unpredictable nature of transportation is exemplified.

AI-driven route optimization, which involves calculating the fastest, most efficient path to a destination, serves as a promising solution for this problem. With access to both current and historical data across the transportation network, AI systems can conduct thorough assessments of all variables affecting the route. By accounting for these external circumstances, routes can be dynamically adjusted, thereby reducing transit times.

If widely implemented, route optimization will play a significant role in saving time, fuel, and money for users.

1.2 Traffic Analysis and Detection

The everyday flow of our traffic often appears to be chaotic and unpredictable. However, beneath the surface, there are hidden patterns that AI systems can help uncover and leverage to improve the smoothness of our transportation networks.

AI systems can analyze this real-time traffic flow through a combination of data collection, processing, and advanced modeling and analysis algorithms. In the data collection phase, data is derived from various sources including traffic cameras, ground sensors, and social media. This vast array of data sources delivers crucial information regarding vehicle movement, speeds, queue lengths, and location, all of which are crucial for in-depth traffic evaluations.

The collected raw data is then processed by algorithms to remove noise and erroneous data. With this processed data, AI can gain valuable insights relating to the nature of our transportation network, enabling accurate congestion detection and traffic predictions. This, in turn, leads to informed decision-making and improved traffic management systems.

1.3 Traffic Signal Management

Traffic flow at busy intersections is rarely smooth. With cars coming from multiple directions and pedestrians waiting at crosswalks, the traffic scene can get messy and difficult to manage. Traffic Signal Management (TSM) aims to address this issue.

TSM utilizes AI to control traffic lights in real-time, altering timings based on external variables such as the number of vehicles waiting at the intersection, the number of pedestrians preparing to cross, and even traffic conditions down the road. Research has already demonstrated that this kind of adaptive signal control can improve 'average performance metrics (travel time, control delay, emissions, and fuel consumption) by 10 percent or more' (Federal Highway Administration n.d.: 1). Through these adjustments, AI systems seek to optimize the timing of traffic lights, prioritizing high-congestion routes to reduce the likelihood of bottlenecks. Overall, TSM can reduce transit time for commuters and increase the efficiency of the transportation network, while substantially cutting down on idle times at intersections.

1.4 Predictive Maintenance

Vehicle breakdowns have always been unpredictable. They often lead to long periods of downtime and expensive repairs, both of which disrupt the flow of daily operations. Fortunately, AI-driven predictive maintenance (PdM) has surfaced as a practical solution for this problem.

PdM employs AI to detect operational anomalies, predict equipment failures, and schedule maintenance checks. By processing and analyzing historical records, AI systems can identify specific patterns and characteristics indicative of an approaching breakdown. In the realm of transportation, PdM can be of great benefit as it reduces the risk of unexpected vehicle failures that could potentially endanger lives. Additionally, it prevents mechanical issues from escalating further which ultimately saves money in the long run.

Research has shown that PdM can result in substantial cost savings, with one study stating that its proposed PdM model reduced system costs by 9.68% compared to state-based policy, 39.07% compared to time-based policy, and 39.56% compared to greedy policy (Feng and Li 2022).

1.5 Autonomous Vehicles

Autonomous Vehicles (AVs) are self-driving cars that rely on AI systems to navigate the roads without humans controlling the wheel. The concept of AVs has been met with a surge of interest in recent years, as it promises to revolutionize the way we move around. Already, many companies in the transportation industry have invested heavily into the development of this technology.

AVs stand to make such a large splash in our world because of their ability to enhance safety and efficiency within transportation. Through the assistance of AI, AVs are expected to take over driving tasks such as parking and changing lanes. Supplied by large amounts of data, these systems can make near-optimal decisions in real-time and adapt well to external conditions. Ultimately, upon widespread implementation, AVs will be able to weave through complex traffic scenarios more reliably than humans.

1.6 Smart Public Transit

Unlike individual vehicles that navigate roadways on their own, public transit involves the management and dispatch of a large fleet. From subways and trains to buses, we see that these systems are complex, with numerous variables impacting their efficiency and reliability. Due to the vast array of moving parts within public transit, AI has arisen as a viable solution for improving operational efficiency.

AI can allocate vehicles and drivers based on real-time data, leading to effective fleet management. With AI predictive models and sophisticated algorithms, public fleets can be dispatched swiftly in response to real-time conditions such as unplanned weather and traffic events.

By analyzing historical traffic patterns, current road conditions, and passenger demand at different times and locations, AI systems enable effective and optimized route planning for public transit. Additionally, AI's adaptability allows it to create and adjust schedules in real time, meeting passenger demand while minimizing transit times.

Incorporating AI into public transportation maximizes efficiency and reduces fuel consumption, thereby lowering greenhouse gas emissions. Furthermore, AI-driven improvement in public transportation will encourage more people to use these services, reducing the need for extensive parking infrastructure. Consequently, excess parking spaces can be repurposed for other uses.

1.7 Detection of Illegal Activity

The rise of AI in transportation will undoubtedly be accompanied by an increase of sensors integrated into vehicle infrastructure. While these sensors will primarily be used to provide AI models with visual data for transportation efficiency, they can also play a critical role in detecting criminal activity. With the help of this visual data, AI can detect abnormal traffic behavior that deviates from normal patterns, such as unexpected stops and irregular transit routes. These anomalies may indicate illegal activities such as smuggling, human trafficking, or illegal transportation. Additionally, through image and video analysis of data from transport hubs (e.g., airports, ports, train stations), AI can spot suspicious behavior or identify individuals on criminal watch lists. Already, Transport for London, a government agency that manages the transport network in London, has utilized AI to monitor the London Underground, generating over 44,000 potential security alerts during the test (Burgess 2024).

AI can also be used to prevent illegal activities. By analyzing historical data, AI models can monitor routes and locations prone to trafficking, thereby predicting potential illegal activities before they occur.

1.8 Ridesharing

As the demand for mobility in our world continues to grow, ridesharing has become an increasingly popular alternative to owning a personal vehicle. Ridesharing refers to the service where a driver is matched to a passenger through a platform, allowing the driver to transport the passenger from one location to another, occasionally with the option to pick up other passengers heading the same direction.

AI holds such large significance in this specific industry due to its ability to analyze historical records of rides and predict location-specific demand. By using this information, companies can identify areas where they need to boost driver availability and save resources in others. Additionally, with consideration for factors such as weather, route closures, and congestion along with proximity and user preferences, AI can intelligently match drivers with passengers. Even in instances where a driver must pick up multiple passengers on the way to a destination, AI-driven technology can plan the best routes to each stop, saving time and fuel.

1.9 Air and Sea Traffic Control

Just as AI is revolutionizing ground transportation, it has also brought remarkable advancement in air and sea transportation, particularly in the realm of air and sea traffic control.

Air traffic control (ATC) directs and streamlines the flow of aircraft traffic both in the airspace during flight and on the ground during takeoff and landing. The operation of ATC relies on a vast range of data in the National Airspace System (NAS) which includes flight plans, aircraft information, radar tracked real-time aircraft positions, and weather data. AI powered algorithms can analyze this NAS data and derive optimal real-time insights, streamlining flight route, predicting and mitigating potential collision hazards, and reducing risks associated with human errors. For example, AI systems can suggest changes in flight paths to avoid congestion and bad weather conditions, thereby reducing delays.

Similarly, AI can also be applied to enhance the safety and efficiency of sea transportation. By analyzing historical data, AI can predict trajectories of ship movements. The pilots can use this valuable information to develop and execute safe navigation plans, especially in congested shipping lanes and ports.

2. AI Technology in Transportation

With its unique ability to tackle complex problems, AI holds enormous potential in the development of our transportation networks. Tremendous progress, made in the past decade, has brought our society closer to the first tremors of a seismic shift in our conceptualization of mobility. However, though AI is poised to play a fundamental role in directing our daily lives, it remains shrouded in mystery to the public eye. Ensuring greater accountability in the integration of AI into transportation is crucial; for this reason, we must aim to demystify the underlying technology.

2.1 Machine Learning

Taking a closer look at a transportation scenario reveals an intricate web of moving parts.

Imagine traffic at an intersection: vehicles of varying speeds converging, pedestrians milling about, and unpredictable drivers cutting into lanes all demonstrate the complexity of transportation.

This large number of dynamic and changing variables presents a problem for the conventional traffic management systems we once relied on. Traditional methods have struggled to adapt to modern issues such as rising mobility demand, greenhouse gas emissions, and increased congestion. Given the consequences of these challenges, identifying a robust solution has never been more critical. Fortunately, Machine Learning (ML) has emerged as a promising solution.

ML, a subfield of Artificial Intelligence (AI), involves developing statistical algorithms that allow computers to perform tasks without explicit instructions. Utilizing vast amounts of data, ML systems are trained to find patterns and make predictions, even when faced with dynamic variables and uncertainties. These capabilities position ML to make significant contributions to transportation.

Among the notable ML techniques is Supervised Learning (SL), which uses labeled data to train models. To understand this concept, imagine presenting a computer with data pairs where inputs and their corresponding outputs are known. Through training, the computer identifies patterns, eventually making accurate predictions for new inputs. SL is particularly valuable in traffic supervision, taking diverse inputs such as time, weather conditions, and road types to predict traffic outcomes. This enables accurate predictions of congestion and road patterns, allowing for improved traffic management systems like signal control and dynamic routing.

Reinforcement Learning (RL) is another ML model, differing from SL by using trial and error to guide behavior. A computer interacts with its environment, learning through a system of rewards and penalties. Over time, it takes actions that maximize rewards, improving its performance. In transportation, RL is commonly used to train autonomous vehicles. As vehicles interact with roads, pedestrians, and other vehicles, RL enables them to adjust based on feedback, enhancing their ability to navigate dynamic traffic scenarios.

Deep Learning (DL), a subset of ML, uses neural networks with multiple layers that transform input data into increasingly abstract representations. This process enables DL to uncover patterns traditional ML might miss. Its ability to process vast amounts of data is particularly advantageous in transportation, where complex relationships must be analyzed. For example, DL can optimize vehicle routes by considering factors like weather, time, and traffic conditions, making efficient navigation more feasible than ever before.

2.2 Computer Vision

Shifting transportation infrastructure toward autonomy necessitates smarter visual systems. A closer look at the latest AI advancements reveals our reliance on visual data. Take traffic incident detection: for these systems to work, they must scan the environment and use visual input to identify incidents. This is where Computer Vision (CV) comes into play.

CV is a key subfield of AI focused on acquiring, analyzing, and understanding images or videos. Essentially, CV mimics the human visual system, enhanced by the speed and precision of computers. It uses sensors like RGB cameras and LiDAR to help vehicles "see" their surroundings. These sensors are increasingly prevalent in transportation, providing the high-dimensional data crucial for autonomous vehicle vision and infrastructure.

After collecting visual data, machines detect and recognize objects using deep learning models. Convolutional Neural Networks (CNNs), a standout architecture, use layered processing to recognize and classify images. Depth estimation and motion analysis then create 3D models to track movement.

Together, these systems enable automated transportation to perceive environments and make real-time decisions. A prime example is Tesla's Autopilot system. Using cameras and sensors, Tesla vehicles process real-time data to detect and classify objects such as vehicles, pedestrians, lane markings, and traffic signs. This information allows navigation through complex traffic scenarios.

2.3 Optimization Techniques

An optimization problem refers to the challenge of finding the best solution from a set of other suitable alternatives to fulfill a specific objective. In the modern era of transportation, numerous optimization problems have appeared across the industry, some of which include traffic signal control, route planning, driver-passenger matching, and public transit scheduling. In all these areas of study, we can see how optimization stands as the common goal. For example, in traffic signal control, optimizing involves determining the best timing for traffic lights so that congestion is minimized. Considering the central role optimization problems play in modern transportation development, let's explore two notable AI-driven techniques that have emerged through research.

2.3.1 Genetic algorithms (GAs)

These algorithms mimic natural selection to find optimal solutions. They first start off with a collection of feasible solutions (chromosomes), gradually evolving them through successive generations based on their "fitness". Additionally, by introducing random changes (mutations) and combining solutions (crossover), GAs can thoroughly explore the range of solutions, thereby increasing the likelihood of identifying an optimal solution.

2.3.2 Ant colony optimization (ACO)

ACO is an optimization technique inspired by the foraging behavior of ants. When an ant travels on a path, it leaves behind pheromones which attract more ants. As a result, favored paths have stronger trails of pheromones, reinforcing their attractiveness. Similarly, in ACO, artificial ants first explore the range of solutions, depositing pheromones along their routes. Over iterations, the pheromone levels of each solution is adjusted according to certain criteria. For example, if a certain route or solution proves to be more efficient than the others, then its level is marked higher. Through this process, the algorithm converges toward the optimal solutions.

The "AI-driven" aspect of these optimization techniques comes from the fact that they autonomously explore solutions, evolving over iterative stages to find optimal solutions. It is this autonomy that demonstrates the significant potential of these algorithms in transforming transportation systems.

2.4 Data Acquisition

In the modern era of technology, data abounds. From social media platforms to smart devices and sensors, the amount of data generated has risen at an unprecedented rate, opening doors for innovation across a multitude of industries. This influx of information holds particular significance in the field of transportation; after all, through the collection and analysis of data, we gain valuable insights into traffic patterns and driver behavior, thereby revealing areas for improvement.

Already, AI has been used to leverage these insights, leading to the development of smarter transportation solutions. However, to be effective, these solutions demand a constant stream of accurate data to flow in from across the transportation network. This is what makes data acquisition such an important aspect in the sector - it ensures that AI systems remain "up to date", adaptable to changing conditions and capable of real-time decision making. As a result, a need for robust data acquisition has surfaced in the industry, calling into focus two key areas: the Internet of Things (IoT) and Big Data.

2.4.1 Internet of things (IoT)

The Internet of Things (IoT) refers to a web of interconnected smart devices that collectively work to gather and share data. Some examples of these devices include vehicle sensors, traffic cameras, and GPS systems. Through this process, large amounts of information can be distributed across the transportation network in real-time, enabling accurate updates on current traffic conditions, congested routes, and infrastructure performance. The key impact of IoT lies in its ability to supply AI systems with the data they need to operate. Consider a traffic jam,

specifically one that stretches for miles. In a scenario like this, AI systems require a comprehensive overview of the entire situation before they can narrow it down to the problem area. Accomplishing this task is no easy feat as data must be gathered from multiple sources across the traffic jam. However, with IoT technology integrated into individual vehicles, the process becomes much more efficient. After all, an increase in smart devices translates to a larger stream of data flowing into AI systems. This ultimately results in a more thorough analysis of the traffic scenario.

2.4.2 Big data

Big data refers to extremely large collections of data. In the context of AI and transportation, it involves the real-time processing and analysis of these vast datasets to extract useful insights. An example of this would be the aggregation of data from thousands of vehicle sensors to monitor and optimize traffic flow. Spurred on by the IoT, Big Data is used by AI models to identify patterns and make predictions, thereby enabling informed, real-time decision-making. By playing such a critical role in the development of AI-driven technologies, Big Data allows for the creation of a smarter, efficient, and safer transportation network.

3. Real-World Implications

The role AI will play in our system of transportation will greatly impact the way we get around. In terms of efficiency, integrated AI systems will assist many modes of transport in various ways. For motor vehicles, an AI could function as a GPS but with greater precision and responsiveness. Unlike traditional GPS systems, AI can act as a personal assistant, pointing out traffic rules, construction zones, and incidents on the road. Rather than simply displaying directions, the AI would behave more like a friend in the passenger seat, offering tips to enhance the driving experience. With real-time pathfinding, it can instantly recalculate routes based on current road conditions, staying several steps ahead of traditional systems. These capabilities extend to other modes of transport, with trains optimizing operations and airplanes actively managing resources to reduce fuel consumption.

AI-powered transportation also has the potential to strengthen connectivity and promote sustainability. Intelligent vehicles can navigate complex obstacles to reach remote locations, connecting communities that were once inaccessible. By doing so, AI reduces the need for infrastructure expansion, preserving natural environments while minimizing costs. This balance between connectivity and ecological conservation highlights AI's ability to link people while protecting the planet.

Safety remains one of the most significant benefits of AI integration. Intelligent systems will act as personal assistants for drivers, pilots, and operators, identifying hazards, assessing risks, and monitoring vehicle conditions. In emergencies, AI can immediately notify emergency services, provide repair assistance through integrated databases, and even offer first-aid instructions. By detecting risks and responding proactively, AI prevents accidents and saves lives across various modes of transport, from roadways to aviation.

Beyond individual vehicles, AI can transform the broader transportation network. AI-driven traffic management systems could coordinate vehicles on highways to prevent congestion, reducing delays and fuel consumption. Public transportation systems like buses and subways could optimize schedules and routes dynamically, adapting to passenger demand in real time. Freight logistics would also benefit, as AI systems streamline supply chain operations by predicting delays and rerouting shipments with precision.

Economic accessibility is another key advantage of AI-powered transportation. Intelligent systems lower costs for passengers and businesses alike by improving efficiency and reducing reliance on human operators. Automated shipping and delivery systems accelerate operations and make goods more affordable. Moreover, by connecting remote regions to urban hubs, AI creates opportunities for employment and economic development in underserved areas, fostering a more inclusive global economy.

AI-powered transportation symbolizes a shift toward a more connected and efficient world. These technologies are not just enhancing mobility but fostering a future where sustainability, safety, and accessibility are prioritized. This marks the beginning of a new era—one that is greener, smarter, and more reliable.

4. Challenges and Ethical Considerations

With all these enhancements in AI for the efficiency and sustainability of transportation comes a higher degree of safety. However, this topic can be controversial for many who fear the usage—or rather, the misusage—of AI. While researchers at tech companies like OpenAI and Google claim to be refining the practice of training safe

AI models, there are valid concerns about the misinterpretation of information by AI; any small mistake on the road, sea, or air could prove fatal.

Mistraining often manifests as algorithmic bias, where an AI model repeatedly favors a particular result or category over others, even with varied inputs. This bias typically stems from flawed initial training data, leading to skewed outputs. In the context of transportation, such errors could compromise decision-making processes, increasing the risk of accidents and raising questions about the safety and reliability of AI-integrated systems.

Safety and reliability are not the only concerns. The integration of AI and smart technology has sparked widespread apprehension about data security and user privacy. According to a recent study, 81% of Americans are concerned about how companies use their collected data, and 71% worry about government use of the same data (McClain et al., 2023). When it comes to AI, 81% believe that increased AI usage would make them uneasy about their privacy, compared to only 62% who think AI could improve their lives (McClain et al., 2023). AI-integrated transportation systems could easily track users and collect extensive data. While this could help monitor driver safety records, it could also extend far beyond its intended purpose, compromising confidentiality. Proprietary models, which operate on secretive code, only heighten concerns about hidden motives, such as exploiting user data without consent.

Another issue is interoperability. The safety of AI-integrated transportation hinges on its ability to process accurate and consistent data across networks of interconnected systems. If faulty data infiltrates these networks, it could lead to widespread errors in decision-making or exacerbate algorithmic bias. These issues are difficult to detect, harder to resolve, and could cause prolonged system backlogs, disrupting economies reliant on efficient transportation. Financial constraints further complicate matters, as training effective and sustainable AI models requires substantial time and resources. Without adequate funding, companies may lack the incentive to pursue AI integration, stalling progress.

AI integration also raises economic challenges, particularly job displacement. Between 1990 and 2017, automation caused the loss of around 400,000 jobs in the U.S., equivalent to 23,500 jobs per year (Semuels, 2020). Extrapolating these figures to AI, particularly in the service sector, could reasonably result in even greater job losses over a similar period. While AI simplifies life for consumers, it creates challenges for those whose work it replaces, further highlighting the need for proactive economic planning.

Despite these challenges, there are measures to address these concerns. For instance, Tesla's smart cars are equipped with features to prevent the mishandling of their Autopilot function, a promising step toward safer AI systems. Rapid advancements in AI could lead to models that are more efficient, less prone to bias, and increasingly accident-proof. Regarding data privacy, government intervention, economic incentives, and strict cybersecurity standards could help prevent corporate misuse of data. Additionally, as old jobs are displaced, new roles in emerging fields like AI development could stimulate the economy and create opportunities for those affected.

These are just initial steps toward mitigating the risks of AI. By addressing such concerns proactively, we can prevent incidents on the road, in the air, or at sea. Considering the vast possibilities of AI-integrated transportation—especially in areas with high stakes for human health and safety—is essential to ensure that we pave the best possible path for its development.

5. Current Policy and Regulations

Combatting the challenges of AI integration, companies have implemented policies to ensure safe usage. Tesla, for instance, prevents misuse of its Autopilot feature by requiring drivers to keep their hands on the wheel during self-driving (Tesla n.d.). This underscores the importance of maintaining human oversight in AI systems, especially in transportation.

Other industries also provide valuable examples of responsible AI use. In healthcare, Merative uses AI for medical diagnoses while adhering to HIPAA guidelines to protect patient data and minimize errors. These AI systems work alongside human practitioners, who ensure decisions are accurate and appropriate. Regular testing and maintenance further enhance reliability (Merative 2022).

In finance, companies like JP Morgan Chase rely on AI for managing transactions and analyzing trends. Human analysts validate or correct AI outputs, creating a feedback loop that reduces bias and ensures compliance with regulations. This iterative process improves AI systems over time.

These strategies highlight key lessons for AI in transportation. Companies like Google's Waymo are at the forefront of integrating safeguards, training AI models in virtual simulations to prepare for diverse traffic scenarios (Waymo 2023). These simulations enable the AI to process vast sensor data, identifying patterns to improve road safety. General Motors incorporates redundant systems to ensure safety even during failures (Abuelsamid 2017). Volvo adapts its AI to individual users through feedback, enhancing both safety and usability.

From these practices, a clear regulatory framework emerges for transportation: ongoing maintenance to ensure reliability, human-supervised feedback loops to mitigate ethical risks, and redundant safety measures to address potential failures. Human validation is particularly vital for eliminating bias and ensuring fairness. Redundant systems, widely used in aerospace, add a crucial layer of reliability. Together, these approaches promote safe and efficient AI deployment.

However, regulating AI alone isn't enough—corporations must also be held accountable.

History shows that some companies prioritize profit over ethics, posing risks when AI is misused. In transportation, such misuse could lead to catastrophic consequences, particularly if cyberattacks compromise critical datasets or systems. Government intervention through enforceable regulations can prevent corporate negligence and ensure ethical AI development.

These foundational steps are essential for a future where AI-powered transportation systems operate safely, efficiently, and ethically, setting the stage for responsible AI integration across industries.

6. Future Directions

AI-driven transportation technologies are still far from widespread implementation, largely due to the overwhelming number of unpredictable variables within transportation networks. This complexity highlights the immense challenges AI seeks to address. Despite these hurdles, AI-centered research in transportation is advancing rapidly. Experts predict that the AI transportation market will grow to $10.3 billion by 2030 (MarketsandMarkets 2017). To realize these exciting prospects, it's essential to address key challenges and refine AI systems to meet future demands.

One major focus is increasing the reliability of AI technologies, such as autonomous vehicles, which must operate safely in all traffic and weather conditions. Handling rare and unpredictable scenarios remains a challenge, as demonstrated by studies showing that 94% of autonomous vehicle accidents are caused by passive actions of other drivers (Wang et al. 2020). Continuous development and testing of AI solutions are essential to mitigate these risks and build safer transportation systems.

Improving perception systems is another critical area for advancing transportation infrastructure. Most AI technologies in this sector rely on visual understanding to interpret and respond to their environments. However, smart cameras, a common tool in transportation, face limitations in detecting and tracking objects under environmental interference (Bai et al. 2021). Future advancements will focus on refining sensors, enhancing object detection, and improving data integration to overcome these challenges.

Large datasets are crucial for training effective and reliable AI systems. Yet, obtaining data that covers all possible scenarios, including rare events, remains a challenge. Incomplete or biased data can lead to errors, particularly in morally complex situations, such as deciding between hitting a cyclist or colliding with another car. These dilemmas require universally accepted moral guidelines for AI decision-making (Frank et al. 2019).

Moving forward, researchers must focus on creating comprehensive datasets, refining algorithms for transparency and efficiency, and establishing ethical and regulatory frameworks to address these challenges. Such efforts will shape the future of AI in transportation, ensuring safer and more reliable systems.

7. Conclusion

For decades, we have imagined a world where personal mobility is accessible, efficient, safe, and environmentally friendly. Today, at the intersection of AI and transportation, the ingenuity of scientists and engineers is bringing that vision closer to reality.

With large strides taken in areas such as traffic analysis, autonomous vehicles, and route optimization, a diverse array of AI tools is being leveraged to design the transportation network of the future. Through these developments, not only will our understanding of mobility evolve, but critical issues regarding safety, environmental sustainability, and efficiency will also be resolved. Unfortunately, AI integration has raised concerns relating to data privacy, ethical decision-making, and job displacement. Tackling these problems

requires a general demystification of AI technologies and implementation of comprehensive regulatory frameworks. It is only through this approach that we can ensure transparency and accountability during large scale integration of AI into transportation. Although other factors such as economic costs and infrastructure limitations must be addressed before this transition, the future of AI in transportation holds great promise. The path in front of us will be one of innovation, collaboration, and perseverance, ultimately leading us toward a drastic reshaping of our transportation ecosystem.

References

Abuelsamid, S. (2017) *To make a truly autonomous car, GM is making some big changes under the bolt's hood*, *Forbes*. Available at: https://www.forbes.com/sites/samabuelsamid/2017/09/11/gm-and-cruise-automation-debut-3rd-gen-automated-bolt-designed-for-production/ (Accessed: 24 July 2024).

Bai, J. *et al.* (2021) "Robust Target Detection and Tracking Algorithm Based on Roadside Radar and Camera", *Sensors*, 21(4), p. 1116. doi:10.3390/s21041116.

Burgess, M. (2024) *London Underground is testing real-time AI surveillance tools to spot crime*, *Wired*. Available at: https://www.wired.com/story/london-underground-ai-surveillance-documents/ (Accessed: 02 July 2024).

Federal Highway Administration (n.d.) *Adaptive signal control technologies*. Available at: https://www.fhwa.dot.gov/innovation/everydaycounts/edc-1/pdf/asct_brochure.pdf (Accessed: 16 July 2024).

Feng, M. and Li, Y. (2022) "Predictive Maintenance Decision Making Based on Reinforcement Learning in Multistage Production Systems", *IEEE Access*, 10, pp. 18910–18921. doi:10.1109/access.2022.3151170.

Frank, D.-A. *et al.* (2019) "Human Decision-making Biases in the Moral Dilemmas of Autonomous Vehicles", *Scientific Reports*, 9(1). doi:10.1038/s41598-019-49411-7.

MarketsandMarkets (n.d.) *Artificial Intelligence in transportation market size, report, 2030*, *MarketsandMarkets*. Available at: https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-in-transportati on-market-261260227.html (Accessed: 19 July 2024).

McClain, C. *et al.* (2023) *How Americans view data privacy*. rep. Pew Research Center. Available at: https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/ (Accessed: 10 July 2024).

Merative (2022) *Achieving the benefits of AI with choice and control*, *Merative*. Available at: https://www.merative.com/content/dam/merative/documents/ebook/achieving-the-benefits- of-ai.pdf (Accessed: 24 July 2024).

Semuels, A. (2020) *Machines and AI are taking over jobs lost to coronavirus*, *Time*. Available at: https://time.com/5876604/machines-jobs-coronavirus/ (Accessed: 10 July 2024).

Tesla (n.d.) *Autopilot and full self-driving capability: Tesla Support*, *Tesla*. Available at: https://www.tesla.com/support/autopilot#:~:text=Before%20enabling%20Autopilot%2C%20the%20driver,your%20hands%20on%20the%20wheel.%22 (Accessed: 23 July 2024).

Texas Department of Transportation (n.d.) *Texas motor vehicle traffic crash facts calendar year 2022*. Available at: https://ftp.txdot.gov/pub/txdot-info/trf/crash_statistics/2022/01.pdf (Accessed: 24 July 2024).

Wang, J. *et al.* (2020) "Safety of Autonomous Vehicles", *Journal of Advanced Transportation*. Edited by F. Bella, 2020, pp. 1–13. doi:10.1155/2020/8867757.

Waymo (2023) *Waymo advances AI research with our multifunctional Waymax Simulator*, *Waymo*. Available at: https://waymo.com/blog/2023/10/waymo-advances-ai-research-with-our-multifunctional-w aymax-simulator/#:~:text=Simulation%20is%20a%20critical%20tool,performance%20of%20the%20Waymo%20Driver (Accessed: 24 July 2024).

Students' AI-generated Images: Impact on Motivation, Learning and, Satisfaction

Cornelia Berg[1], Liv Omsén[1], Henrik Hansson[1] and Peter Mozelius[2]
[1]Stockholm University, Sweden
[2]University of Gävle and Mid Sweden University, Sweden

henrik.hansson@dsv.su.se

Abstract: In the contemporary society where the rapid development of generative AI (GenAI) infiltrate daily life, it is imperative for schools to keep up with the development. Future generations are expected to have GenAI skills and knowledge in the same way that traditional literacy is a recruitment condition today. School curricula need to integrate this new technology to support students' learning and development. Research on artificial in education (AIED) has reported on the challenges with the involvement of GenAI in teaching and learning, but also on GenAi as a study support. To not fall behind many schools have started out with AIED initiatives which creates a need for studying how GenAI could be applied in a useful way in teaching and learning activities. The research question that guided this study was: In what ways could the use of GAI in visual form support students' learning process and motivation in upper secondary school settings? The overall research strategy was a qualitative case study approach with investigator triangulation. Data were collected in a combination of observations at workshop sessions and semi-structured interviews. with teachers as well as students. In a six-step inductive thematic analysis data excerpts were coded, aggregated to categories, and presented. Findings indicate that GenAI tools for image generation, can have a positive effect on learning. At the same time that memorisation of information is supported, there were also a positive impact on motivation and student satisfaction. The involvement of image generation tools not a substitute, but rather a complement to traditional teaching and learning activities. The conclusion is that the use of AI in education can offer new learning opportunities, and with the increased use of GenAI, it is crucial for both students and teachers to keep pace. However, this would require that more time and resources for teacher professional development on AIED.

Keywords: Generative AI, GenAI, AIED, AI-generated images, Midjourney, Secondary education

1. Introduction

The rapid development of generative AI (GenAI) reaches many areas where education is one of them. Important to keep up with the introduction of new tools and techniques and to look at the opportunities that open up for an enhancement of teaching and learning activities. As highlighted by Chiu (2023), GenAI tools such as ChatGPT and Midjourney have a potential to make a strong impact on several aspects of education. Rather than seeing GenAI as only a threat, the traditional education system could be reinforced by introducing GenAI tools as study companions (Zarris & Sozos, 2023; Farrelly & Baker, 2023), and that the threshold for students to use GenAI is low (Sullivan et al., 2024). There are also studies discussing the potential of using AI as a teacher assistant (Humble & Mozelius, 2019; Parab, 2020), and as pointed out in the study by Zarris and Sozos (2023), GenAI tools can be both study companions and teacher assistants. GenAI involves systems that generate content based on multi-layered neural networks that can identify and imitate patterns, styles and structures that has been a part of the systems training data (Farrelly & Baker, 2023).

Many recent research studies have focused on how GenAI trained on large language models generates textual output (Barreiro et al., 2024), and the benefits of GenAI in language learning (Roe, 2024; Thorne, 2024). As argued by Law (2024), many studies have highlighted the potential of GenAI chatbots as personalised language learning partners. However, GenAI is a multimodal technique where the output could be text, sound, music images or video (Rabowsky, 2023; Huang, 2023). In parallel with the release of a wide variety of AI tools with different modal outputs there is a need for research on how they could be applied in a useful way in teaching and learning activities. As suggested by Johnson (2023), GenAI tools could be a way of accelerating learning, but with the risk of misleading learners with wrong information. During the last years many image generating tools have been released such as DALL·E 3, Midjourney, Stable Diffusion and Adobe Firefly. Midjourney is one of the image generating tools that has been evaluated to have both a high output quality and a user-friendly interface (jie et al., 2023; Turgay et al., 2023).

Midjourney is also one of the tools that several times has been part of some recent studies on GenAI in educational contexts (Chen et al., 2024), and it was the chosen tool for this study. The aim of this study was to explore the impact of involving GenAI tools in upper secondary school activities and to analyse how AI supported image generation might stimulate motivation and creativity. The main research question to answer was: "In what ways could the use of GenAI in visual form support students' learning process and motivation in upper secondary school settings?"

2. GenAI in educational contexts

GenAI can potentially act as a study companion for future generations. With the help of various AI technologies, complex concepts can be explained, advanced questions answered, and new study techniques based on course content created (Farrelly & Baker, 2023). As society becomes digitalised, it is important to keep students motivated and provide them with the skills to navigate in an innovative society (Byundyugova et al., 2022). At the same time, research highlights the importance of creative thinking in school (OECD, 2019a; 2019b) and how visual literacy increases motivation to learn (Kaya, 2019). Technologies are evolving and becoming more complex, which may result in visual communication playing a greater role (Kaya, 2019). Furthermore, in the PISA 2021 *Creative Thinking Framework* (2019b) PISA highlights the importance of training students in their creative abilities. A student trained in creative thinking may find it easier to adapt to the future we are facing and deal with more complex challenges and find innovative solutions both locally and globally.

A digitalized society also places demands on the school that students and teachers possess the skills required to keep up with the development of society. The school should help a student to feel that they are part of today's society and that their resources can contribute to development (OECD 2019b). Furthermore, the school is facing an educational renewal and change in the school's operating culture. Research shows that students prefer technological platforms to physical books and teachers prefer platforms such as Google Classroom and Moodle (Chen, Xiw & Hwang, 2020). The development of intelligent technology is changing the whole school structure, operation and governance (Chen, Xiw & Hwang, 2020). It is becoming more difficult to determine what is reality and what is AI and therefore society must collectively find new ways of working, learning and studying as GenAI will become a major part of our everyday lives (Farrelly & Baker, 2023).

AIED (AI applications in education) involves various tools and applications with AI for students and teachers whose purpose is to support current teaching (Chen, Xiw & Hwang, 2020). In Sweden, artificial intelligence will be a new subject in upper secondary school and KomVux (Municipal adult education) from autumn 2024. The aim is to teach AI from a societal perspective and how AI can be used for problem solving. The Swedish National Agency for Education states that teachers are interested in how technology can be integrated into current teaching. New technological tools, including GenAI technology, are introduced in schools and affect how students think, work and students' learning ability. Examples of GenAI tools are Midjourney which generates images based on the user's textual description (Yanbo & Chuanlan, 2024) and ChatGPT which has become a major part of the school just in the last year. Despite a great deal of criticism towards the tools, some argue that it also brings positive aspects. Moreover, this type of technology will be part of our everyday lives whether we like it or not. If used correctly, GenAI technology can be revolutionary and accelerate our learning (Johnson, 2023).

Visual creation in school environments creates motivation in students. Pisa believes that motivation is a key in creative thinking and it can be measured in the engagement a student feels towards a task or lesson (OECD, 2019b). An engaged classroom is often a result of so-called active learning which means that students are involved and challenged to reflect on what they are doing and why (Bonwell & Eison, 1991). Furthermore, the classroom should be an innovative, permissive and inspiring environment that challenges students in their creative thinking, cognitive knowledge and openness to new ideas (OECD, 2019b). Previous research emphasises that allowing students to be creative and independent when working on tasks also contributes to a supportive environment during lessons (Bonwell & Eison, 1991). The relationship between creativity and education can be seen as natural and historically creativity has been important in schools (Smith & Smith, 2010). Creative learning involves techniques teachers use to help students learn new material that they can later apply to new problems. Finally, creative learning is about how students use strategies to process new information (Mayer, 1989).

3. Method and sample

The used research strategy was a qualitative case study approach with multiple data sources analysed by investigator triangulation. Data were collected in a combination of observations at workshop sessions and semi-structured interviews. with teachers as well as students. A case study is recommended when there is one phenomenon, or a few phenomena, that should be investigated in depth, with the aim of establishing a deeper understanding for events, experiences, relations and processes (Denscombe, 2014). Furthermore, case studies are often used to explore and discover a new information field, which also was the purpose with this study.

3.1 Sample and data collection

Observations involved nine students, aged 17 to 18 years, who were enrolled for the second year at an upper secondary school in central Stockholm. Students were observed when writing prompts and generating images in the GenAI tool Midjourney. After the observations four of the students answered questions in a semi-

structured interview about their perceptions of the image generation. Moreover, a semi-structured interview was conducted with the teacher in the video conference system Zoom since there was no time to do the interview the same day as the student interviews were conducted at the school. All the interviews were recorded and later transcribed by two of the authors. Interviews were conducted in Swedish.

The observations were conducted in a workshop setting where students were working in pairs, but with their individual computers. Three images should be generated and discussed, where the first one was generated without any instructions to the students. For the second image they were encouraged to enhance the prompting, and then to compare the results. Finally, for the third image the instructions were instructed to formulate a prompt for an image with relations to a societal or historical theme.



Image 1. Midjourney generated images of Winston Churchill when he was elected Prime Minister

3.2 Data analysis

The collected data were analysed in a six-step inductive thematic analysis as outlined by Braun and Clarke (2006). A thematic analysis is a frequently used method for identification and analysis of patterns and correlations in a data set. This was used in this study both for the interviews and the observations. The aggregate data were analysed and grouped into categories that have a potential to answer the research question. According to Braun and Clarke the first step is about familiarising yourself with the data, which started out in the transcription of interviews and the editing of notes. In the next step code excerpts that had a potential to contribute to the answering of the research question were marked for further analysis. This was followed up by a step where the identified patterns and codes were aggregated to subcategories and preliminary categories (themes). Then the preliminary codes were revised and checked for consistency before the final naming of the categories. Finally, a narrative presentation of the found categories were created with the idea of explaining the relevance of the categories and their relationships.

3.3 Ethical considerations

This study has followed the ethical guidelines from the Swedish Research Council (Vetenskapsrådet, 2017, p.40). The study has strived to keep all involved as anonymous as possible and all personal information about teachers and students were removed early in the process, or never entered, to eliminate the possibility of connecting any images or answers to specific individuals. Moreover, all participation in the observation was conducted on a voluntary basis, with the right of withdrawal without any explanation at any stage of the process. According to the guidelines of informed consent, all informants have signed a form with information about the study. An important aspect for this case study was to consider the specific ethical guidelines for upper secondary school students. According to Integritetsskyddsmyndigheten / The Swedish Authority for Privacy Protection (IMY, 2024), students over 16 years of age have a certain legal capacity (rättshandlingsförmåga), and could in general give their consent without signatures of the parents or guardians.

4. Results and discussions

Observations in classroom sessions and follow-up interviews provided the themes or categories of: 'Active learning', 'User-friendliness,' 'Motivation and learning outcomes', 'Critical thinking', and 'Integration in classroom activities'. The found themes are presented and discussed one by one here below.

4.1 Active learning

Students found it very motivating to actively create their own AI images, it was expressed by a student in this way: "*You are the one who creates and it will be more fun than sitting and looking at other pictures*." The oral classroom discussions based on prompts and images further discloses students actual understanding and expands it. Another process, which can reinforce learning, is the iterative process towards a more accurate image; writing additional prompts to improve the accuracy of the image. This refining process is supporting the learning process; when writing a prompt - viewing an image, critically assessing it and rewriting the prompt to get a better result - the learning increases. *"You can do it together and discuss, we create a picture together, absolutely a good learning technique, find five errors and correct them"* (Teacher). A method which could be used in education is comparative analysis of images, comparing students' images with real and authentic illustrations.

There are important learning aspects which may be less tangible to assess as a teacher, but nevertheless are important for the learner. One student expressed that he acquired a deeper and more empathic understanding of war conditions when viewing images. A subject which traditionally has been taught by using texts and statistics. *"We have worked a lot with history and if you see a picture of what it was like in the trenches, for example, you get a completely different perspective on what it was actually like. And it may not be part of how I will answer the question, but it gives me a better idea of how it actually was."* Teachers saw significant potential in using AI tools like Midjourney to foster collaborative discussions and enhance understanding of historical and literary contexts. The ability to generate visual aids was seen as a powerful complement to traditional teaching methods.

4.2 User-friendliness

All students agreed that Midjourney was easy to learn and use. "I thought it was easy to start writing and it was easy to understand how the program works, especially when you got to test it once because then you saw how it reacts to what is written in the prompt." The quick adoption and understanding of how to use the system by students differ from the teacheŕs view, who stated time constraints as a major limit for exploring and learning new tools and approaches. It is a bit strange that learning professionals, i.e teachers, find it harder to learn new things than their students. The teacher's opinion was that AI takes time to learn, but that it on the other hand accelerates learning and thereby saves time. What is the definition of 'Education'? Well, there are many but one is 'accelerated learning'.

Most things can be learned without formal education, but the advantage with education is that with the structure, content, mentoring etc provided in education, the learning journey is faster. AI can greatly increase learning speed in almost all subjects. The teacher in this study is probably more open minded than most teachers, some definitely against digitalisation. "There is a trend in school where there are those who feel that we should go back to students having paper and pencil, not only writing on screen but also by hand, because it provides more learning. I think it is a complement and there must be both, the school will become more digitised and teachers think it's fun". The primary challenge that remains is the limited time available for teachers to familiarise themselves with new AI tools and to train students in their use. This necessitates a strategic approach to professional development and resource allocation within schools.

4.3 Motivation and learning outcomes

All students agreed that images improved learning. "It becomes easier to write and understand instead of remembering texts. I find it easier to remember if I link an image to a text. If I were to write a text and get a picture, I could remember the picture and connect the words I wrote to create it to it and then remember better." There is a risk of learning wrong based on incorrect pictures, but even a biased picture could be useful as a trigger for classroom discussions, analysing what is right with this picture? What is wrong with this picture? In such reflections, using student generated pictures with teacher guided analysis in discussions students can move from rough and superficial understanding to deeper and more critical understanding.

In fact, analysis of the image is also an analysis of the prompt that student wrote to generate the image, which is an indication of the current state of his/her understanding. Moreover, students claimed that Midjourney facilitated learning by making it easier to memorise information and understand complex concepts. The visual

nature of the tool was particularly effective in enhancing engagement and motivation. The effectiveness of Midjourney varied depending on the input data. While it successfully generated realistic images for certain historical events, it struggled with others, highlighting the importance of accurate and relevant input data. Results also indicated that students developed visual literacy skills, which are crucial for interpreting and analysing AI-generated images. This ability is seen as equally important as traditional literacy skills.

The use of Midjourney GenAI images in this study released creativity and intense activity by the students who immediately embraced the technology. There is a tension between creativity and control of the learning processes, on the one hand open mind, think out of the box, generating new ideas and on the other hand controlling what is true, what to be trusted, how to judge, how to assess and award grades. Obviously, students like the creativity aspect which is engaging, new, student centred and motivating. From the teacher's point of view engaged and motivated students are positive aspects of course, but there are concerns about how to control the learning process, the accuracy of results, how to keep abreast with the new technology and be ahead of students. Changing teaching and learning activities is not only about time, it is also about willingness to change.

4.4 Critical thinking

Several students remarked that the AI generated images are not real authentic illustrations, therefore, in order to acquire deep and correct knowledge other sources of information are needed. They also stressed the importance of writing correct and relevant prompts in order to get a useful and trustworthy illustration of the topic. Knowing how to write efficient and correct information is a key skill which is not provided by the AI tool, but rather still a teacher´s/advanced learner´s skill. *"One should not rely entirely on AI, so perhaps not as the primary study tool. You also have to use other information to understand that it is not only what is shown that is the truth, because it is still a fake picture."*

In order to write a good prompt, one has to know a lot already, the importance of pre-knowledge and the ability to critically assess the GenA image result based on previous knowledge is emphasised. *"You have to know a lot of things to use it in a good way. For example, you have to make a house, or like this task a house in Jon Utzon style, now she got a house that matched well, but then she had given more information, and looked at lots of pictures and knows what it should look like".*

Students found that Midjourney facilitated learning by making it easier to memorise information and understand complex concepts. The visual nature of the tool was particularly effective in enhancing engagement and motivation. Despite these benefits, there were concerns about the potential for AI to distort reality. Students and teachers emphasised the need for critical thinking skills to accompany the use of AI-generated content. Basic skills in the field of AI generated images can also be a way to reinforce critical thinking in the emerging field of "deepfakes" (Mutillo-Ligorred et al., 2023).

4.5 Integration in classroom activities

Gen AI for image generation could be used for individual learning activities, memorisation, classroom discussion and as a basis for process examination as well as final examinations. The teacher suggested that prompt texts could be a basis for grading. *"A student can show through a prompt so I as a teacher can see what he/she has changed in the picture, it could be assessment material if that is what you are looking for".* Actually, there are reasons to believe that how to use AI will be a general lifelong skill required in future work life and therefore should not be banned in education but rather encouraged. Students turn to the Internet to learn about how to write prompts as a Pro, since the teacher has limited knowledge. *"If you have classes where you learn how to write prompts and learn the basics and you do that then it's bound to be good. [...] I found a guy on TikTok who was a prompt pro so he was a pro at using AI and for me it was a bit weird that there are already those and then I thought that's probably a very good thing to learn and invest in it"*

The teacher had several innovative ideas of how to use GenAI image generating for student activities in different subjects: in literature students could generate images illustrating the novel they read, in history they could illustrate correctly the time period - 16th century London compared with 19th century London - for example, and in architecture/art they could illustrate houses built in a special style. The teaches explains a student activity including GenA image creation and being able to argue and defend that the image is correct:

> *"Let's say it would be an assessment situation, then maybe I could ask the students to create a certain scene in history and when the student feels that he/she have created a scene that describes it, for example the Second World War, or it could be [more specific] the Battle of Lutzen and then the student has to argue why he/she think this picture is a good answer, in this way we could work"*

While the students mainly raised positive aspects of using GenAI, the teacher was more reserved. Time and organisational constraints, a dystopic future and….. was mentioned by the teacher. *"We don't really have the time to continue our education, it goes so fast, the school is behind in that way, here you have to freelance as an individual teacher and just decide that you think it's fun to try, even though you don't master it", and that "Schools haven't really caught up, there are still teachers who have learning tasks made by AI, and we are facing some kind of tipping point and schools need to give time for teachers to sit and learn the systems"*

> *"Students are ahead of us, it's hard to keep up. But as teachers, you contribute with critical thinking about how they use AI, but it requires that teachers are given space to further train in this. Now the new curriculum will also be added and will mean that teachers must learn them and change the subjects how we should think, the time that is available in school will now be used for that, and then there is not much time for AI".*

5. Conclusion

In conclusion, AI image generating tools like Midjourney can enhance student learning by making learning more student centred, interactive, fun and engaging. However, critical analysis and open discussions with the teacher about the validity of images generated by students are important in order to arrive at correct understanding and conclusions. Successful integration of such technologies requires addressing the time constraints faced by teachers and ensuring that both students and teachers are equipped with the necessary skills to use these tools creatively, critically and effectively. The integration of AI in education should be accompanied by efforts to enhance students' critical thinking skills. While AI offers numerous benefits, it should complement rather than replace traditional teaching methods. Human judgement and interaction remain essential components of effective education. Future research should focus on expanding the sample size and exploring long-term impacts of AI integration in various educational contexts to validate and generalise these findings.

References

Barreiro, A., Lloret, E. and Gerazov, B., 2024. A linguistic approach to Natural Language Generation: The language spectrum, language models, and OpenLogos. Open Research Europe, 4(123).

Bonwell, C., Eison., J, A., 1991. Active Learning: Creating Excitement in the Classroom. ASHE-ERIC Higher Education Reports.

Braun, V. and Clarke, V., 2006. Using thematic analysis in psychology. Qualitative research in psychology, 3(2), pp.77-101.

Byundyugova, T., Babikova, A., Kornienko, E., 2022. Development of Educational Motivation of Adults with the Help of Visual Technologies. International Journal of Media and Information Literacy, 7(1), pp.28-37.

Chen, Y., Zhang, X. and Hu, L., 2024. A progressive prompt-based image-generative AI approach to promoting students' achievement and perceptions in learning ancient Chinese poetry. Educational Technology & Society, 27(2), pp.284-305.

Chen, X., Xie, H., Hwang, G-J., 2020. A multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers, Computers and Education: Artificial Intelligence, Volume 1.

Chiu, T.K., 2023. The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. Interactive Learning Environments, pp.1-17.

Denscombe, M., 2014. The Good Research Guide : For Small-scale Research Projects (5th ed.). Berkshire: Open University Press.

Farrelly, T. and Baker, N., 2023. Generative artificial intelligence: Implications and considerations for higher education practice. Education Sciences, 13(11).

Huang, Y., 2023. The future of generative ai: How genai would change human-computer co-creation in the next 10 to 15 years. In Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play, pp. 322-325.

Humble, N. and Mozelius, P., 2019. Teacher-supported AI or AI-supported teachers. In the European Conference on the Impact of Artificial Intelligence and Robotics -ECIAIR 2019, pp. 157-164.

IMY 2024. Integritetsskyddsmyndigheten/The Swedish Authority for Privacy Protection, Retrieved (16/04/2024) from: https://www.imy.se/en/

Jie, P., Shan, X. and Chung, J., 2023. Comparative Analysis of AI Painting Using [Midjourney] and [Stable Diffusion]-A Case Study on Character Drawing. International Journal of Advanced Culture Technology, 11(2), pp.403-408.

Johnson, W.L., 2023. How to harness generative AI to accelerate human learning. International Journal of Artificial Intelligence in Education, pp.1-5.

Johnson, Lewis.W., 2023. How to Harness Generative AI to Accelerate Human Learning. International Journal of Artificial Intelligence in Education.

Kaya, Mustafa., 2019. The impact of Visual Literacy Awareness Education on Verbal and Writing Skills of Middle School Students. International Journal of Education & Literacy studies, [8].

Law, L., 2024. Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. Computers and Education Open.

Mayer, Richard E. 1989. Cognitive Views of Creativity: Creative Teaching for Creative Learning. Contemporary Educational Psychology, 14, 203-211.

Mutillo-Ligorred, V., Covaleda, I., Ramos-Vallecillo, N. and Fayos, L., 2023. Knowledge, integration and scope of deepfakes in arts education: the development of critical thinking in postgraduate students in primary education and master's degree in secondary education (No. ART-2023-135204).

OECD., 2019a. PISA 2018 Assessment and Analytical Framework, PISA, OECD Publishing, Paris.

OECD., 2019b. PISA 2021 creative thinking framework (third draft), OECD.

Parab, A.K., 2020. Artificial intelligence in education: teacher and teacher assistant improve learning process. International Journal for Research in Applied Science and Engineering Technology, 8(11), pp.608-612.

Rabowsky, B., 2023. Applications of generative ai to media. SMPTE Motion Imaging Journal, 132(8), pp.53-57.

Roe, J., 2024. Exploring the Impacts of GenAI on English and Applied Linguistics: Implications for a Future-Ready Journal. Journal of English and Applied Linguistics, 3(1), p.3.

Sullivan, M., McAuley, M., Degiorgio, D. and McLaughlan, P., 2024. Improving students' generative AI literacy: A single workshop can improve confidence and understanding. Journal of Applied Learning and Teaching, 7(2).

Smith, Jeffery K. & Smith, Lisa F. (2010). Educational Creativity. Cambridge University Press.

Thorne, S.L., 2024. Generative artificial intelligence, co-evolution, and language education. Modern Language Journal, 108(2), pp.567-572.

Turgay, O., Kaptan, M., Öktem Erkartal, P., Demir, T. and Doğukanlı, A.Ç., 2023. Interdisciplinary trajectories: regenerating Barcelona Pavilion with midjourney. Interior Architectural Issues Design Theory Philosophy, Livre de Lyon, pp 299-320.

Zhang, Y., Liu, C., 2024. Unlocking the Potential of Artificial Intelligence in Fashion Design and E-Commerce Applications: The Case of Midjourney. J. Theor. Appl. Electron. Commer. Res. pp 664-670.

Zarris, D. and Sozos, S., 2023. Educational Artificial Intelligent Chatbot: Teacher Assistant & Study Buddy. Master Programme in Data Science. Luleå University of Technology Department of Computer Science, Electrical and Space Engineering.

Vetenskapsrådet/The Swedish Research Council. 2017. God forskningssed [Good research practice]. Retrieved (15/04/2024) from: https://www.vr.se/download/18.2412c5311624176023d25b05/1555332112063/God-forskningssed_VR_2017.pdf

Artificial Intelligence, Smart Topological Data Analysis and Chaos in Business Continuity Management: *The Case of COVID-19 in Birmingham Airport*

Carlos Gonçalves and Carlos Rouco
Lusofona University, Intrepid Lab, Lisbon, Portugal
CETRAD, Porto, Portugal

p6186@ulusofona.pt
carlos.rouco@ulusofona.pt

Abstract: The latest state-of-the-art empirical methods from chaos theory have incorporated smart topological data analysis (STDA) combining chaos theory methods, topological machine learning, adaptive artificial intelligence systems, topological data analysis and fractal analysis methods for attractor reconstruction and the topological study of the dynamics, with impact on risk science and complexity research. In the current work, we apply a topological adaptive AI system to the study of Birmingham airport's air traffic dynamics, and employ topological data analysis, chaos theory methods and multifractal analysis to research the resulting dynamics. Our results show the presence of a form of stochastic chaos with a low-dimensional attractor associated with a long wave dynamics in the pre-COVID-19 period, which continues in the COVID-19 crisis and subsequent recovering around a rising trend, with the topological AI system able to adapt to the COVID-19 crisis and predict with high performance the dynamics during this period. Multifractal analysis methods, applied to the adaptive topological AI system's residuals, show that the dynamical noise affecting the chaotic attractor is multifractal, with a multifractal phase transition occurring during the COVID-19 recovery period. Implications of the methods and results for business continuity management are drawn.

Keywords: Business continuity management; VUCA; BANI; Topological AI; Smart Topological Data Analysis; Chaos Theory; Risk

1. Introduction

A new organizational context is characterized by the terms volatility, uncertainty, complexity, and ambiguity (VUCA) (Bennett and Lemoine, 2014; Sinha and Sinha, 2020). The most accurate descriptions are quick and chaotic changes, lack of standards, or the constant outdating of plans and projects (Nowacka and Rzemienniak, 2022). The VUCA environment strategically challenges organizations and individuals to achieve success (Gaule et al., 2023). VUCA clearly reflects the characteristics of the emerging world through the aspects of volatility, uncertainty, complexity, and ambiguity, as well as by responses to performing in such an environment of vision, understanding, courage, clarity, adaptability, and agility (Fuentealba et al., 2023; Gaule et al., 2023). Also, the terms brittle, anxious, nonlinear, and incomprehensible (BANI) signifies the fragility of systems, processes, or organizations to sudden disruptions, emphasizing the need for resilience (Kaplan and Mikes, 2016; Linnenluecke et al., 2020).

VUCA and BANI are essential concepts in business continuity management (BCM) and resilience (Galaitsi et al., 2023). VUCA emphasizes the need for organizations to adapt to complex environments, requiring agility and flexibility in emergency response strategies. BANI opposes rigid bureaucratic structures, which can hinder adaptive decision-making in unpredictable situations (Bennett and Lemoine, 2014). To effectively address these challenges, organizations must adopt a multifaceted approach, implementing dynamic and resilient organizational structures that enable swift decision-making and coordination while acknowledging the complex nature of crises (Hrynchak and Motuzka, 2023).

In a VUCA and BANI context, organizations need to be always prepared for the various events that may partially or totally affect their operations. For this preparation, organizations need to incorporate and keep up-to-date with new technologies in their multiple processes, such as artificial intelligence. Like this, to support business continuity management, artificial intelligence (AI) can play a strategic role in organizations, as indicated.

AI enhances risk identification and assessment and management by analysing vast data across operations, supply chain, cybersecurity, natural disasters, enabling organizations to develop effective mitigation strategies and enhance resilience (Fortunato et al., 2023; Grimwade, 2023). By recognizing patterns and trends, AI can help predict potential disruptions before they occur, allowing businesses to take proactive measures to mitigate risks (Cebulla, 2023). AI aids in data-driven decision making, enhancing strategic decision-making during crises and enhancing operational continuity by integrating BCM practices with vast dataset analysis (Xiangwei et al., 2022). AI can assist decision-makers during crisis situations by providing timely insights and recommendations based

on the analysis of relevant data (Yue & Shu, 2024). AI-driven predictive analytics and scenario planning enhance business resilience by simulating scenarios and predicting disruptions, thereby enhancing organizations' contingency plans and strategies (Consilvio et al., 2024). By analysing different contingency plans and their effectiveness, organizations can better prepare for different eventualities and refine their BCM strategies accordingly (Machado et al., 2022). AI-driven analytics can enhance organizations' resilience by continuously assessing and adapting their BCM practices, based on past incidents and response efforts.

Generally speaking, AI can significantly improve business continuity management by offering advanced analytics (Tyagi, 2022), real-time monitoring, decision support, automation, and predictive capabilities, enabling organizations to anticipate, prepare plans, respond, and recover from operational disruptions. Also, AI improve organizational resilience in several keyways, as indicated: Organizational resilience encompasses various dimensions, including operational, Effective leadership, communication, innovation, financial, strategic, technological, supply chain (Kassa, 2023), and human resource resilience, all of which contribute to an organization's ability to withstand and recover from disruptions (Al-Banna et al., 2022). Overall, by integrating AI technologies with BCM practices and organizational resilience strategies, organizations can enhance their ability to anticipate, respond to, and recover from disruptions, ensuring continuity of operations and minimizing the impact on stakeholders (Soldatos, 2021).

Chaos theory, enhanced by new methods from Smart Topological Data Analysis (STDA) that integrate topological machine learning, adaptive AI and topological data analysis tools, offers powerful new technological tools for BCM, including the ability to find patterns in random-looking data, making sense of complex topological patterns in time series, identifying possible exogenous and endogenous risk sources for organizations, adapting to crises and providing solutions for dealing with dynamical changes in critical business variables, without having to specify a model, but rather, instead, uncovering the topological regularities in the temporal data for characterising that data, prediction and scenario analysis.

In the present article, we address the application of these new STDA and chaos theory's empirical methods to airport traffic dynamics prediction. We apply these methods to Birmingham airport's total instrument flight rules (IFR) movements series from 2016 to 2023, finding the presence of a form of stochastic chaos characterised by a low-dimensional attractor associated with a long wave dynamics affected by multifractal noise in the period from 2016 to 2019.

We, then, address the impact of the COVID-19 crisis on the air traffic dynamics. A critical finding is that the topological adaptive AI system is capable of predicting the series during the crisis and the subsequent recovery period with a higher performance than in the pre-COVID-19 period, we also uncover a phase transition in the multifractal spectrum of the noise affecting the chaotic dynamics during the crisis period.

In section 2, we review the main concepts and methods, in section 3, we present the main results, in section 4, we provide for an overall discussion of the relevance of the applied methods to BCM and in section 5 we present the main conclusions of the work around the main empirical results for the Birmingham airport and address the relevance of the results for future research.

2. Methods

Chaos theory initially dealt with bounded nonperiodic dynamics with sensitive dependence upon initial conditions in low-dimensional deterministic systems, leading to random-looking trajectories in nonlinear deterministic systems with a small number of degrees of freedom. The theory then grew to address both networks of nonlinear oscillators (Kaneko and Tsuda, 2001), which are high-dimensional systems, as well as systems with external dynamical noise, that is, systems that exhibit chaos in the deterministic nonlinear dynamics but are also affected by noise, this is the research field of stochastic chaos (Kapitaniak, 1990), which is of particular relevance for risk science and for dealing with VUCA since it deals with endogenous and exogenous sources of risk, including the dynamical impact of noise leading to topological changes in the system's dynamics.

The identification of chaos when one only has available a time series $y(t)$, with $t = 1,2,\ldots, T$, involves the delay embedding of this series and the subsequent research of the properties of the embedded trajectory, with the phase space trajectory described by a point $p(t) = (x(t-(d-1)h),\ldots, x(t-2h), x(t-h), x(t))$ in a d-dimensional space (Takens, 1980; Harte, 2001; Gonçalves, 2022, 2023a,b).

In general, the choice of the lag depends upon analysing the series' memory, the choice can be based on the first zero crossing of the autocorrelation function, the first zero crossing of the partial autocorrelation function or the first minimum of the mutual information. For power law chaos that is, chaos with a power law spectrum, that has a very slow decay in the autocorrelation function, the first minimum of the mutual information can lead

to a too short a lag and the first zero crossing of the autocorrelation function can lead to too long a lag, due to the power law scaling associated with the chaotic attractor. In this context, the partial autocorrelation function often works better at providing for a lag reference (Gonçalves, 2023b).

For the choice of dimension and subsequent analysis of the attractor based on the application of topological machine learning, one needs to find the dimension for which the topological regularities allow for a better prediction of the target series, it is here that adaptive AI equipped with topological machine learning enters into play as a part of topological data analysis methods, this is the basis of STDA and has proven effective when dealing with emergent noisy chaotic attractors in complex systems (Gonçalves, 2022; 2023a; 2023b). This method was first introduced in (Gonçalves, 2022) to deal with chaos in the COVID-19 epidemiological series, in the case where a bifurcation occurred in the Oceania region and was subsequently employed in hospitalizations from COVID-19 (Gonçalves, 2023a) and financial volatility analysis (Gonçalves, 2023b).

The method is based on selecting the embedding dimension, from a set of dimensions, that leads to the best performance, for an adaptive artificial agent equipped with a topological machine learning unit and a sliding window for learning, in the prediction the next value of the target series from the previous phase point, in this way, one is effectively searching for the embedding that leads to the most exploitable topological order that allows the prediction of the target series (Gonçalves, 2023a,b).

The use of a topological machine learning unit is key here, since one is trying to find the embedding, from a set of embeddings, that has the strongest link between the topological structure of the reconstructed attractor and the target series' dynamics, allowing the application of further topological data analysis methods to characterise the attractor and study the main dynamics in conjunction with the topological machine learning results.

Since we will be working with a *k*-nearest neighbours' graph analysis, the machine learning algorithm that we work with for the dimension selection is the *k*-nearest neighbours' algorithm, using a sliding window for relearning and employing a grid search to both calibrate the value of *k* and to select the optimal embedding dimension from a set of dimensions. In this case, for each value of *k*, we extract the dimension that minimizes the relative error measured by the root mean squared error divided by the series' amplitude (RMSE/Amplitude). The relative error metric is useful since it allows for a comparison of the performance between the two periods under analysis (pre-COVID-19 and the COVID-19 and subsequent recovery period). The sliding window for relearning allows for the adaptation to attractor epochs as well as to possible bifurcations (Gonçalves, 2022).

After calibrating the value of *k* and the optimal dimension, we also apply a *k*-nearest neighbours' graph analysis, fixing the dimension to that obtained for the optimal pair (*k*,*d*) and calculate, for different values of *k*, the graph's Kolmogorov-Sinai entropy (K-S entropy) and the relative degree entropy, which is the Shannon entropy of the graph's degree distribution divided by the maximum entropy, in order to characterise the topological properties of the dynamics.

We then analyse the behaviour of the RMSE/Amplitude for the adaptive topological agent and the two entropy measures obtained for each *k*-nearest neighbours' graph, to find the relation between the predictability and the two graph entropy measures. In this case, we find that the increase in entropy is related to a decrease in the relative error, which means that there are topological patterns associated with the nearest neighbours' connections in the reconstructed attractor that allow for prediction.

After performing this topological analysis we report, for the optimal pair (*k*,*d*), the correlation between the adaptive topological agent's predictions and the observed values as well as the explained variance, which allows us to characterise the level of noise.

In the next step, since the nearest neighbours' algorithm can be used as noise filter, we use the agent's predictions as a noise-reduced series and estimate the Lyapunov spectrum on the resulting signal using Eckmann *et al.*'s (1986) method in order to search for markers of chaos, namely, a positive largest Lyapunov exponent, we also report the Kaplan-Yorke dimension (Kaplan and Yorke, 1979; Frederickson *et al.*, 1983), which allows for a characterisation of the noise-reduced reconstructed attractor's dimensionality. In this case, we find evidence of chaos in the airport traffic's long wave dynamics, which is captured by the adaptive topological agent.

After this analysis we turn to the noise process itself, calculating the residuals and employing multifractal detrended fluctuation analysis (MFDFA) to characterise the noise process in terms of possible multifractal signatures which is characteristic of complex turbulent dynamics (Gorjão et al., 2022).

In a multifractal process, given the *q*-th order moments of the variations of a variable measured at different lags *s*, the logarithm of the moments is a function of the logarithm of the lag *s* multiplied by $1+\tau(q)$, where $\tau(q)$ is the

scaling function equal to $qh(q) - 1$, where $h(q)$ is the generalized Hurst exponent. While monofractal scaling is characterised by a single Hurst exponent, leading to a fixed scaling law, in multifractal scaling the Hurst exponents are a function of the order $h(q)$ (Mandelbrot, 1997).

Multifractals cannot be modelled by traditional linear time series analysis methods, which do not capture the scaling involved as addressed in Mandelbrot (1997), such scaling is associated with complex nonlinear processes not corresponding to standard independent and identically distributed noise. Stochastic chaos with multifractal scaling noise is a more complex process then an independent and identically distributed (IID) noise and a monofractal noise process.

It turns out that stochastic chaos with multifractal noise is the type of stochastic chaos that characterises the air traffic data, furthermore, we will find that there is a significant change in the multifractal scaling from the pre-COVID-19 to the COVID-19 and subsequent recovery period, with impact in risk analysis. This change is characterised by a symmetry transformation of the spectrum, which corresponds to a multifractal phase transition, a particularly relevant point when addressing the impact of the crisis on the air traffic dynamics.

Now, for the period from 2020 to 2023, we analyse the ability of the topological adaptive agent to predict the target series when the disruption from COVID-19 occurs and the subsequent process of recovery using the previously obtained embedding.

The advantage of using a machine learning method with a sliding window for relearning is that it allows for an artificial agent to adapt to a change in a process, capturing both a trend and fluctuations around this trend, without having to apply any transformation to the series. The main point, in our case, is to evaluate the impact of the crisis on the ability of an adaptive topological agent to exploit the topological information in the data, using the previously reconstructed attractor, and to adapt to the crisis period. It turns out that the predictability of the target series using the previous embedding leads to an increased and high prediction performance of the adaptive topological agent during the COVID-19 period. Which means that the exploitability of topological information for prediction increased during the COVID-19 period.

After analysing the change in performance, since in the COVID-19 period we find a linear trend and long wave dynamics around this trend which are captured by the topological adaptive agent, we detrend the agent's predictions and once more estimate the Lyapunov spectrum, finding that the long wave dynamics still shows evidence of chaos.

We also calculate the residuals and apply multifractal analysis to these residuals, allowing us to characterise the dynamics pre-crisis and during the crisis period both in terms of the main chaotic dynamics as well as in terms of the associated noise patterns.

3. Findings

In Figure 1, we show the daily Birmingham airport's total IFR movements from 2016-01-01 to 2023-12-31. As can be seen, in the pre-COVID-19 period (2016-01-01 to 2019-12-31), there is a long wave dynamics around a stationary mean with evidence of turbulence, in the period from 2020-01-01, there is a large drop in traffic due to the COVID-19 crisis and a subsequent recovery in the form of a linear increasing trend with the long wave dynamics occurring around that linear increasing trend.

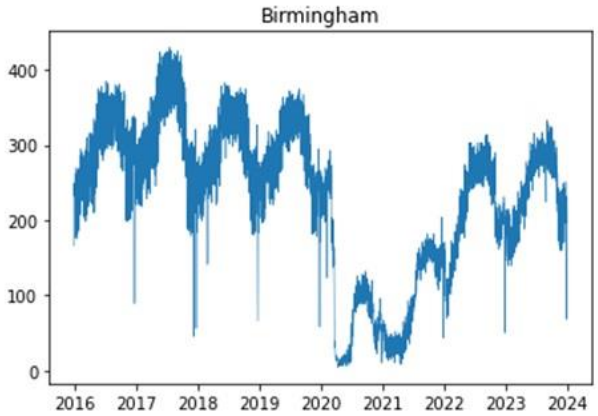


Figure 1: Daily Birmingham airport's IFR movements from 2016-01-01 to 2023-12-31.

Considering, first, the pre-COVID-19 period, in the spectral analysis (figure 2), we find that there is a power law decay associated with long range dependence and a few significant spikes which can be associated with periodic signatures, such a spectrum is consistent with a chaotic dynamics with power law scaling in the spectrum near a periodic window (onset of chaos) (Gonçalves, 2022).

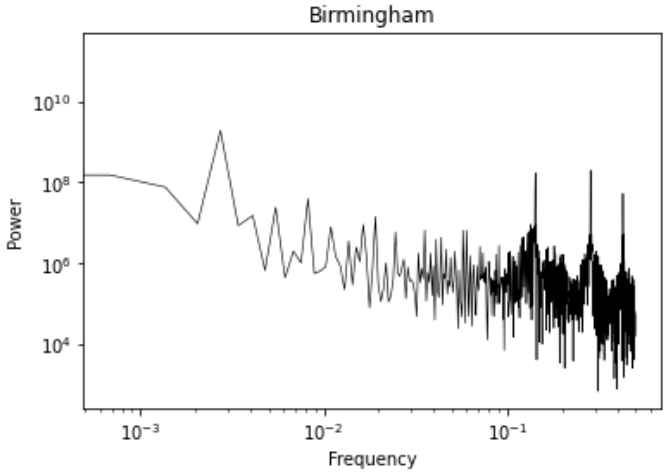


Figure 2: Power spectrum for the period from 2016-01-01 to 2019-12-31.

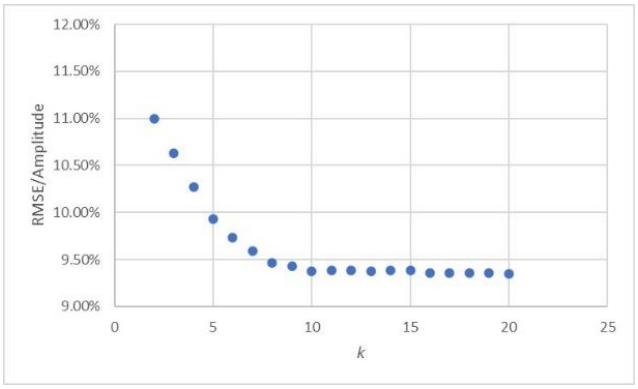


Figure 3: RMSE/Amplitude in percentage for the adaptive topological learner as a function of *k*, for the period from 2016-01-01 to 2019-12-31.

In order to select the lag for phase space reconstruction, due to the long-range dependence, we use the first zero crossing of the partial autocorrelation function, which leads to a lag of 8 days. To select the embedding dimension, in terms of exploitable topological information, we follow the previous section's described methodology and deploy an adaptive topological learner, with a *k*-nearest neighbours' learning unit, distance-based weights, KD-tree learning, sliding learning window of 30 days. We extract for different values of *k* the dimension with the lowest ratio of RMSE divided by the data amplitude. In this case, we find that, for a set of 2 to 20 dimensions, and for *k* ranging from 2 to 20, the optimal dimension is always 4, which provides strong evidence favourable a low-dimensional attractor being present in the airport's dynamics.

As shown in Figure 3, the relative error decreases with the number of nearest neighbours, converging to values below 9.4%, with the lowest value being obtained for k = 20, leading to an RMSE of 36.0028 which corresponds to an RMSE/Amplitude ratio of 9.3514%, that is, the RMSE represents 9.3514% of the total data amplitude. The correlation between the observed and predicted series is 0.7603 and the explained variance is 57.59%, which indicates the presence of noise.

The decrease in the relative error with *k* is directly related to the increase in both the degree entropy and K-S entropy of the k-nearest neighbours' graph for the reconstructed four-dimensional attractor, both these entropy measures grow with the logarithm of the number of nearest neighbours, as shown in Figure 4, which means that the rise in the topological complexity of the *k*-nearest neighbours graph with *k*, measured by the graph's entropy measures for the reconstructed attractor, leads to a rise in the predictability of the target series.

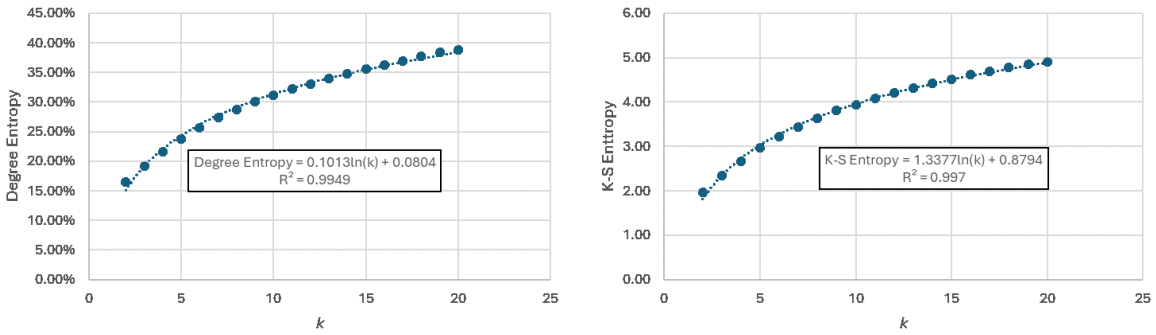


Figure 4: K-nearest neighbours graph's entropy measures as a function of the number of nearest neighbours (*k*), obtained from the embedded series, for the period from 2016-01-01 to 2019-12-31.

As shown in Figure 5 (top left), the topological learner is capable of capturing the long wave dynamics, the Lyapunov spectrum (figure 6 (right)), calculated for the predicted values (figure 6 (left)) which effectively function as a filter, shows a convergence of the spectrum with the largest Lyapunov exponent being positive and equal to 0.0017 which indicates a dynamics that is close to the onset of chaos, for the long wave dynamics. The Kaplan-Yorke dimension estimated from the Lyapunov spectrum is 1.1016 which reinforces the evidence of a chaotic attractor associated with the long wave dynamics.

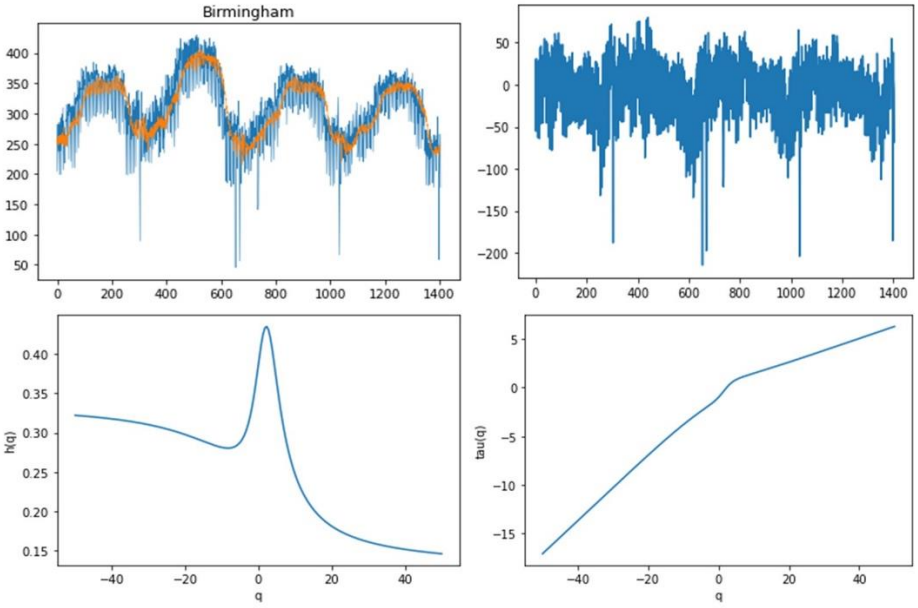


Figure 5: Observed (blue) versus predicted (orange) series for Birmingham for the period from 2016-01-01 to 2019-12-31 (top left), residual series (top right); generalized Hurst exponents' distribution for the residuals series with *q* from -50 to 50 divided by steps of 1/200, 500 lags, from 1 to 1.7, with degree 2 polynomial fit (bottom left); multifractal scaling function (bottom right).

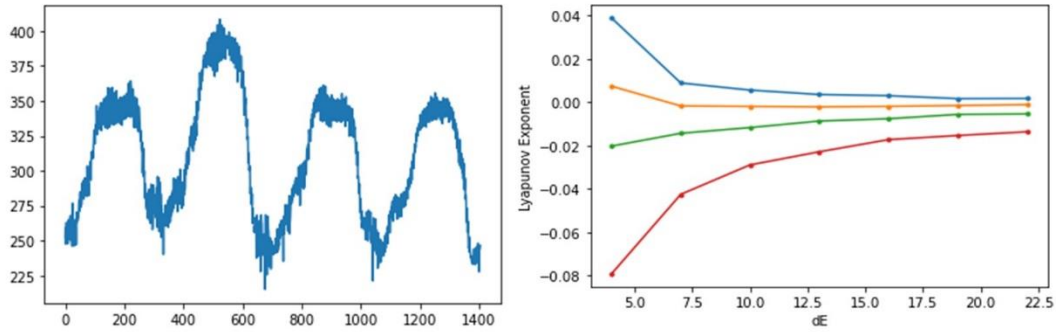


Figure 6: AI-predicted series for Birmingham for the period from 2016-01-01 to 2019-12-31 (left) and estimated Lyapunov spectrum (right) for the four-dimensional embedding of the predicted series.

The noise, in turn, which is estimated from the residuals series (Figure 5, top right) shows evidence of multifractal scaling with a wide spectrum of exponents (Figure 5, bottom left) and a scaling function that increases with the moment order, in each case a degree 2 polynomial fit was used. The estimated generalized Hurst exponents range from 0.1460 to 0.4350, with an estimated spectrum amplitude of 0.2891 for moment orders ranging from -50 to 50. The evidence is, thus, favourable to a stochastic chaotic dynamics, with a low-dimensional attractor associated with long wave dynamics, near the onset of chaos, with the stochastic component being associated with a multifractal noise process.

Now, during the COVID-19 period and subsequent recovery, the airport traffic dynamics changed, there is a large drop in the airport traffic followed by a recovery trend with the large waves occurring around the trend (Figure 7, top left).

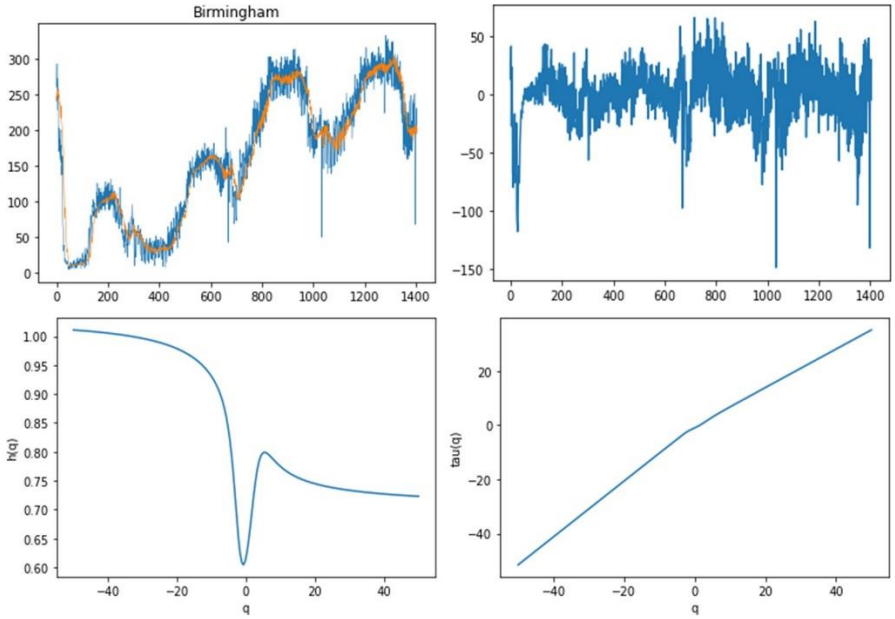


Figure 7: Observed (blue) versus predicted (orange) series for Birmingham for the period from 2020-01-01 to 2023-12-31 (top left), residual series (top right); generalized Hurst exponents' distribution for the residuals series with *q* from -50 to 50 divided by steps of 1/200, 500 lags, from 1 to 1.7, with degree 2 polynomial fit (bottom left); multifractal scaling function (bottom right).

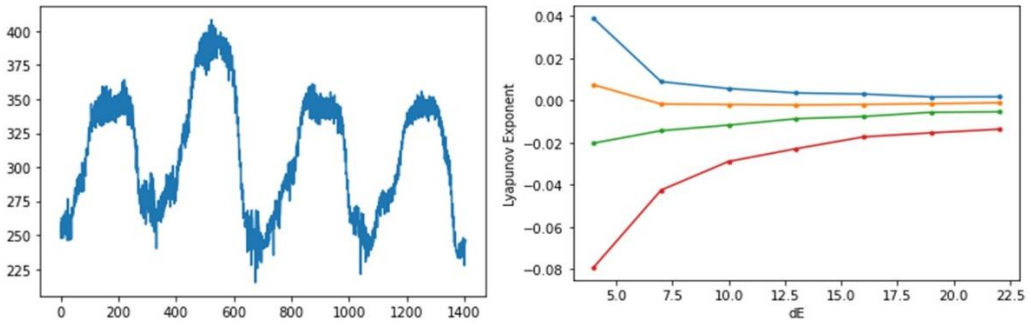


Figure 8: AI-predicted series for Birmingham for the detrended predicted series (left) for the period from 2020-01-01 to 2023-12-31 after dropping the first 100 observations in Figure 7, and estimated Lyapunov spectrum (right) for the four-dimensional embedding of the detrended predicted series.

It turns out that the adaptive topological agent, operating with the embedding parameters for the pre-COVID-19 period is able to adapt to the trend and predict the series during the crisis period with a drop in the relative error (RMSE/Amplitude) from the previous pre-COVID-19 period value of 9.3514% to 7.56%, a rise in the correlation between the observed and predicted values from 0.7603 to 0.9624 and a rise in the explained variance from 57.59% to 92.59%, which is a significant rise in predictability.

There is also a change in the noise process along the predicted trend, in this case, the noise, shown in Figure 7 (top right) is still multifractal, but there is a multifractal phase transition with the spectrum rising to higher values

for the generalized Hurst exponents (Figure 7 bottom left) and there is a 180º rotation with respect to the pre-COVID-19 period, the symmetry transformation is a reflection. The scaling function still rises with the order and the estimated exponents range from 0.6045 to 1.0108, with an amplitude of 0.4063. In this way, there took place a multifractal phase transition characterised by a translation, a reflection and a dilation of the previous multifractal spectrum.

If we drop the first 100 data points which match the air traffic from the COVID-19 lockdown period we get the recovery trend, removing that linear trend we find the presence of the long wave dynamics during this recovery period (figure 8, left). Estimating the Lyapunov spectrum for the series using the previous embedding parameters, we still find a convergence of the spectrum with a positive largest Lyapunov exponent, which is indicative that the long wave dynamics along the rising recovery trend is still one of chaos, however, the largest Lyapunov exponent increased to 0.0028 and the Kaplan-Yorke dimension increased to 1.3690, which means that, while we still have a dynamics near the onset of chaos, there is a slight rise in the largest Lyapunov exponent (Figure 8, right).

4. Discussion

Chaos theory has been applied in aeronautics to address problems as diverse as air traffic flow, turbulence modelling, aircraft engineering as well as aircraft accidents (Lan, 2003; Yu and Li, 2019; Zhang *et al.*, 2020; Zhang *et al.*, 2022). In the current article, we dealt with the COVID-19 crisis impact on air traffic for the Birmingham airport, showing how STDA can be effectively applied to research the pre-crisis dynamics as well as the impact of the crisis and subsequent recovery.

Combining adaptive topological AI, chaos theory and multifractal analysis methods allows for STDA to become effective when dealing with stochastic chaos in complex systems, where emergent chaotic attractors are affected by complex noise processes (Gonçalves, 2022, 2023a,b). In this context, topological AI is a fundamental part of the analysis for embedding selection and to study the attractor's main properties when used in combination with other topological data analysis methods. STDA is robust enough to deal with bifurcations and change in attractor stability as well as the impact of large disruptive events in a system's dynamics, this later point is well-illustrated in the current article.

The insights gained from the application of STDA and chaos theory to deal with complex systems' dynamics involving business processes offers a way to gain strategic intelligence on the systems' dynamics and to reduce the strategic surprise on large events as well to gain insights for scenario analysis and to plan responses to future crises, becoming an important methodological basis for data-driven BCM.

In this context, the current work shows that topological AI is able not only to capture major properties of emergent chaotic attractors but also to adapt to disruptions and changes in dynamics associated with large impact events that may disrupt business processes offering new ground for risk science methods, theories and solutions to support BCM and in dealing with VUCA, offering a data-driven ground on which to address volatility, uncertainty, complexity, and ambiguity.

The capturing of turbulence and complex volatility patterns, the ability to reduce uncertainty especially in the context of crises, the ability to find patterns in complex dynamics and the insights offered by chaos theory's new STDA-based methods provide for a basis for understanding complex systems and reducing ambiguity through data-grounded scientific knowledge of complex systems' dynamics. In this way, the integration of topological machine learning and adaptive AI in topological data analysis methods within the context of chaos theory offers an important basis for BCM's processes, technologies and methods.

5. Conclusions

We applied smart topological data analysis (STDA) methods combining adaptive topological AI, topological data analysis, chaos theory and multifractal analysis, to study Birmingham airport's traffic dynamics and the impact of the COVID-19 crisis on that dynamics. Finding evidence of a form of stochastic chaos, characterised by a long-wave chaotic low-dimensional attractor affected by multifractal dynamical noise.

During the COVID-19 pandemic and subsequent recovery period, we found that the adaptive topological agent is capable of adapting to the crisis and predicting the subsequent dynamics. Several factors contribute to this, on the one hand, there is the predictable rising trend, on the other hand, the increase in the largest Lyapunov exponent is small and the largest Lyapunov exponent, while positive, is still, characteristic of a chaotic dynamics near the onset of chaos, the higher persistence in the noise process may have also contributed to this dynamics,

because it leads to longer laminar periods in the turbulence which may have contributed to a decrease in the error.

In this way, rather than decreasing the predictability, during the crisis period and subsequent recovery the airport traffic actually became more predictable with the topological order being exploitable with the previously estimated embedding. Another important point for complexity research is that we uncovered a phase transition in the multifractal noise process during the COVID-19 crisis recovery process, characterised by a multifractal spectrum translation, dilation and reflection. The importance of the methods and findings for BCM were also addressed in the discussions' section.

The methodology and results of the present work can be extended to other airports, with impact for airport management, including the prediction of air traffic using artificially intelligent systems with adaptive learning windows and risk analysis with implications for scenario analysis, decision making technologies, data-driven business BCM, crisis management and resilience.

References

Al-Banna, A., Franzoi, R.E., Menezes, B.C., Rogers, S., and Kelly, J.D. (2022). "Roadmap to digital supply chain resilience". Computer Aided Chemical Engineering, 49, 571–576. DOI 10.1016/B978-0-323-85159-6.50095

Bennet, N., and Lemoine, J. (2014). "What VUCA really means for you". Havard Business Review, 92 (1/2)

Cebulla, A., Szpak, Z., and Knight, G. (2023). "Preparing to work with artificial intelligence: assessing WHS when using AI in the workplace". International Journal of Workplace Health Management, 16(4), 294–312. DOI 10.1108/IJWHM-09-2022-0141

Consilvio, A., Vignola, G., López Arévalo, P., Borinato, M., and Crovetto, C. (2024). "A data-driven prioritisation framework to mitigate maintenance impact on passengers during metro line operation". European Transport Research Review, 16(1), 6. DOI 10.1186/s12544-023-00631-z.

Eckmann, J.P., Kamphorst, S.O., Ruelle, D., and Ciliberto, S. (1986). "Liapunov exponents from time series". Phys Rev A Gen Phys, 34(6):4971- 4976

Fortunato, H.E., Marcomini, K., Soares, G.J., Reis, F.J., and Abbas, C. (2023). "Optimizing Industrial Risk Management Through AI and No-Code Platform: A Global Solution for the Oil and Gas Industry". Society of Petroleum Engineers - ADIPEC, ADIP 2023. DOI 10.2118/216426-MS.

Frederickson, P., Kaplan, J., Yorke, E., and Yorke, J. (1983). "The Lyapunov Dimension of Strange Attractors". J. Diff. Eqs., 49(2): 185–207.

Fuenteable, D., Flores-Fernández, C., and Carrasco, R. (2023). "VUCA bibliometric and content analysis". Revista Espanola de Documentacion Cientifica, 46(2), e354.

Galaitsi, S. E., Pinigina, E., Jeffrey, M., Pescaroli, G., Keennan, J., and Linkov, I. (2023). "Business continuity management, operational resilience, and organizational resilience: Commonalities, distinctions, and synthesis". International Journal of Disaster Risk Science, 14(5), 713–721. DOI 10.1007/s13753-023-00494-x

Gaule, E., Jovarauskiene, D., Petrauskiene, R., Pravalinskas, M., and Rauleckas, R. (2023). "Managerial approaches, frameworks, and practices for business model application in public services management in the VUCA environment". Engineering Management in Production and Services, 15(3), 84-100. doi: 10.2478/emj-2023-0022.

Gonçalves, C.P. (2022). "Low Dimensional Chaotic Attractors in SARS-CoV-2's Regional Epidemiological Data". Int J Swarm Evol Comput, 11(9): 1000271.

Gonçalves, C.P. (2023a). "Low Dimensional Chaotic Attractors in Daily Hospital Occupancy from COVID-19 in the USA and Canada". Int J Swarm Evol Comput, 12(1):1000291.

Gonçalves, C.P. (2023b). "Chaos-Induced Self-Organized Criticality in Stock Market Volatility-An Application of Smart Topological Data Analysis". Int J Swarm Evol Comput, 12:331.

Gorjão, L.R., Hassana, G., Kurths, J., and Witthauta, D. (2022). "MFDFA: Efficient multifractal detrended fluctuation analysis in python". Comp Phys Commun, 273:108254.

Grimwade, M. (2023). "The potential impacts of the digital revolution on the operational risk profiles of banks". Journal of Risk Management in Financial Institutions, 17(1), 71–88.

Harte D. (2001). Multifractals – Theory and Applications. Chapman & Hall/CRC.

Hrynchak, N., and Motuzka, O. (2023). "The key administrative competencies of managers required for company development in the bani world". Economics - Innovative and Economics Research Journal, 11(1), 289-305. doi: 10.2478/eoik-2023-0012.

Kaneko, K, and Tsuda, I. (2001). Complex systems: chaos and beyond: chaos and beyond: a constructive approach with applications in life sciences. Springer Science & Business Media.

Kapitaniak, T. (1990). Chaos in systems with noise. World Scientific.

Kaplan, J., and Yorke, J. (1979). "Chaotic behavior of multidimensional difference equations" (PDF). In Peitgen, Walther, H. O. (eds.). Functional Differential Equations and the Approximation of Fixed Points. Lecture Notes in Mathematics. Vol. 730. Springer. pp. 204–227.

Kassa, A., Kitaw, D., Stache, U., Beshah, B., and Degefu, G. (2023). "Artificial intelligence techniques for enhancing supply chain resilience: A systematic literature review, holistic framework, and future research". Computers and Industrial Engineering, 186, 109714. DOI 10.1016/j.cie.2023.109714.

Lan, L.W., Lin, F.-Y., and Kuo, A.Y. (2003). "Testing and Prediction of Traffic Flow Dynamics with Chaos", Journal of the Eastern Asia Society for Transportation Studies, Vol 5, pp 1975-1990.

Machado, R.C., Rizzi, J., Xavier, C., Chaves, V., and Oliveira, V. (2022). "Development of drill pipes failure prediction models and operational management system using real-time data analytics and AI". Proceedings of the IADC/SPE Asia Pacific Drilling Technology Conference, APDT, 2022-August. DOI 10.2118/209858-MS.

Mandelbrot B.B. (1997). "Fractals and Scaling in Finance – Discontinuity, Concentration, Risk". Springer.

Nowacka, A., and Rzemieniak, M. (2022). "The impact of the VUCA environment on the digital competences of managers in the power industry". Energies, 15, 185. https://doi.org/10.3390/en150101.

Sinha, D., and Sinha, S. (2020). "Managing in a VUCA World: Possibilities and pitfalls". Journal of Technology Management for Growing Economies, 11(1), 17–21. https://doi.org/10.15415/jtmge.2020.111003

Soldatos, J., Philpot, J., and Giunta, G. (2021). "Cyber-physical threat intelligence for critical infrastructures security: Securing critical infrastructures in air transport, water, gas, healthcare, finance and industry". Cyber-Physical Threat Intelligence for Critical Infrastructures Security: Securing Critical Infrastructures in Air Transport, Water, Gas, Healthcare, Finance and Industry, 1–443. DOI 10.1561/9781680836875.

Takens F. (1980). Detecting strange attractors in turbulence. Dynamical systems and turbulence, Springer, 366-381.

Tyagi, D.M., and Pandita, D. (2022). "Artificial intelligence and people analytics-A key to employee engagement". 2022 International Conference on Sustainable Islamic Business and Finance, SIBF 2022, 224–228. DOI 10.1109/SIBF56821.2022.9940118.

Xiangwei, K., Ziming, W., Mingzheng, W., and Xiangpei, H. (2022). "Trustworthy decision-making in artificial intelligence-enabled systems: Progress and challenges". Journal of Industrial Engineering and Engineering Management, 36(6), 1–14. DOI 10.13587/j.cnki.jieem.2022.06.001.

Yue, Y., and Shyu, J.Z. (2024). "A paradigm shift in crisis management: The nexus of AGI-driven intelligence fusion networks and blockchain trustworthiness". Journal of Contingencies and Crisis Management, 32(1), e12541. DOI 10.1111/1468-5973.12541.

Yu, H., and Li, X. (2019) "On the chaos analysis and prediction of aircraft accidents based on multi-timescales". Physica A, 534, 120828.

Zhang , Z., Zhang, A., Sun, C., Xiang, S., and Li, S. (2020). "Data-Driven Analysis of the Chaotic Characteristics of Air Traffic Flow". Journal of Advanced Transportation, 2020, 8830731.

Zhang, Z., Cui, Z., Wang, Z., and Meng, L. (2022). "Research on Chaotic Characteristics and Short-Term Prediction of En-Route Traffic Flow Using Ads-B Data". Fractals, https://doi.org/10.1142/S0218348X2340131X.