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.

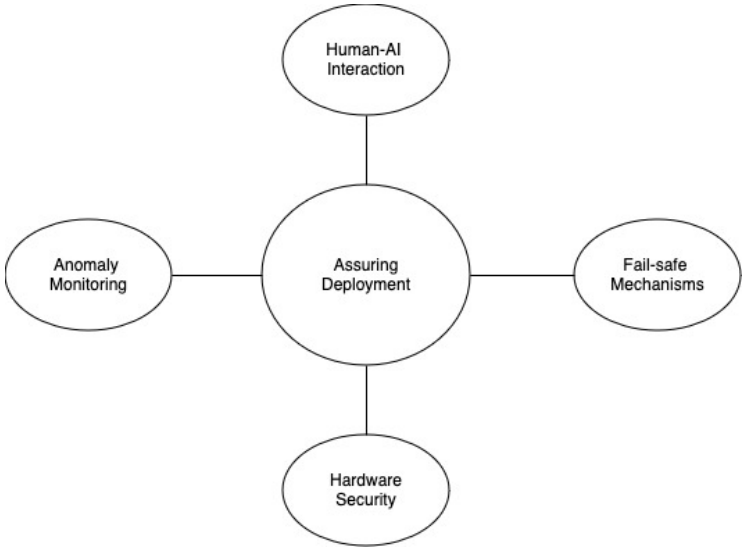

Figure 1: A taxonomy for assuring deployment of an AI system

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.

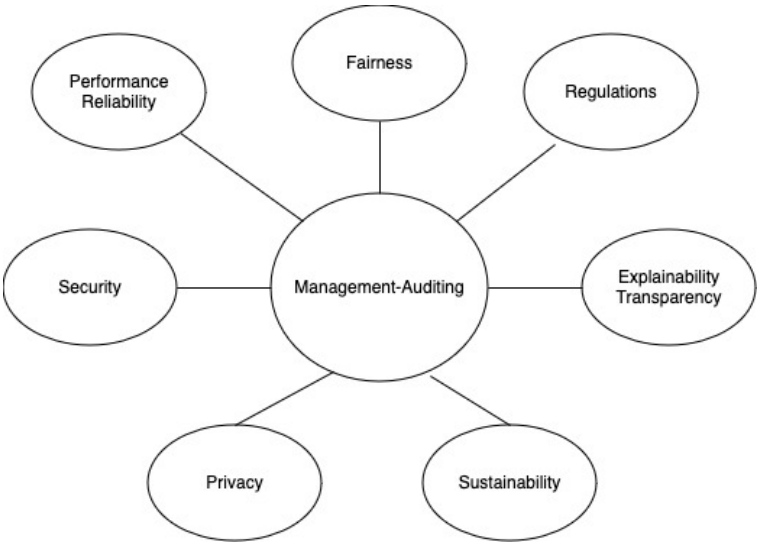

Figure 2: A taxonomy for auditing management of an AI system

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.

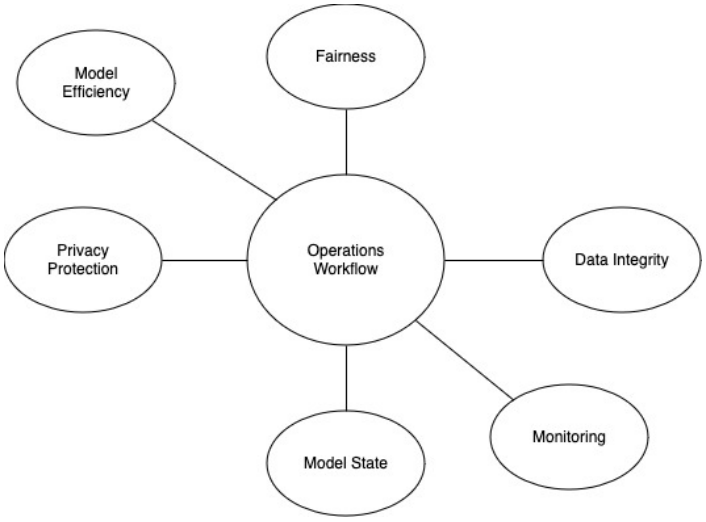

Figure 3: A taxonomy for operations workflow in AI systems

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*.