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.