REFLECTION: ETHICS IN COMPUTING IN THE AGE OF GENERATIVE AI

Maria Ingold
12693772
Unit 1
Research Methods and Professional Practice
University of Essex Online
27 May 2024

CONTENTS

INTRODUCTION 3
DESCRIPTION 3
ANALYSIS 4
EVALUATION 6
CONCLUSION 7
REFERENCES 8

INTRODUCTION

The Artificial Intelligence (AI) field has existed since 1956, however, the impact of AI significantly increased in 2023, with OpenAI's ChatGPT hitting 1 billion visits in February 2023, only three months after launch (Carr, 2024; Russell & Norvig, 2021). Although responsible AI practices predate 2023, this generative AI explosion has prompted further ethical evaluation.

While there is no global regulation, Corrêa et al. (2023) reviewed 200 ethical AI guidelines to establish global consensus across 17 groups of principles. Their paper evaluates 2014 to 2022, including the "AI ethics boom" of 2017 to 2019, but does not reflect the generative AI impact from 2023 onwards.

DESCRIPTION

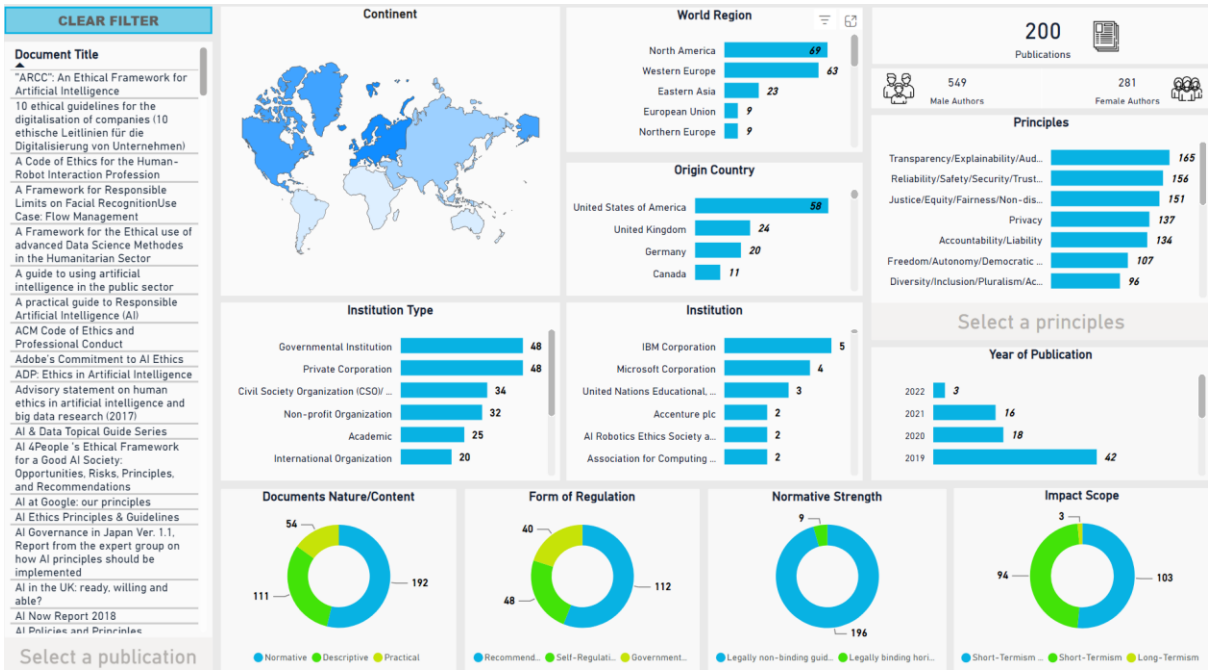

FIGURE 1 | https://nkluge-correa.github.io/worldwide_AI-ethics/dashboard.html

Corrêa et al.'s (2023) GitHub code repository facilitates reproducibility and extensibility of their analysis, while their Microsoft Power BI dashboard (Figure 1)

exemplifies their key points globally, and enables filtering, for instance, by continent, document, or principle.

TABLE 1 | Study by numbers

Study	Publications	Years	Principles
Corrêa et al. (2023)	200	2014-2022	17 principles
Hagendorff (2020)	22	2016-2019	22 issues
Fjeld et al. (2020)	36	2016-2019	47 principles in 8 themes
Jobin et al. (2019)	84	2011-2019	11 principles

Fjeld et al. (2020) selected their 36 documents for prominence and variety, while Corrêa et al. (2023) aim to minimise potential small sample size bias by selecting 200 publications, but still underrepresent regions like Africa and China.

Furthermore, Corrêa et al. (2023) address the unwieldiness of Fjeld et al.'s (2020) 47 principles by reducing to 17, reflecting the need to arrive at global high-level (rather than detail) consensus, while maintaining more granularity than eight themes. Unfortunately, the authors misidentified Hagendorff as 21, not 22, principles.

ANALYSIS

TABLE 2 | Top Five Principles

Study	Top 5 Principles
Corrêa et al. (2023) (Global)	• Transparency / Explainability / Auditability (83%) • Reliability / Safety / Security / Trustworthiness (78%) • Justice / Equity / Fairness / Non-discrimination (76%) • Privacy (69%) • Accountability / Liability (67%)
Hagendorff (2020)	• Privacy Protection (82%) • Fairness, Non-discrimination, Justice (82%) • Accountability (77%) • Transparency, Openness (73%) • Safety, Cybersecurity (73%)
Fjeld et al. (2020) (By theme)	• Fairness and Non-discrimination (100%) • Privacy (97%) • Accountability (97%) • Transparency / Explainability (94%) • Safety / Security (81%)
Jobin et al. (2019)	• Transparency (87%)

	• Justice and Fairness (81%) • Non-maleficence (71%) • Responsibility (71%) • Privacy (56%)

Presented together, Table 2 demonstrates similarities between global studies, even when low sample. Corrêa et al. (2023) and the prior studies generally overlap in all five areas (Jobin et al.'s (2019) non-maleficence includes safety), which would seem to establish their objective of a general global consensus.

However, priorities continue to change over time, as seen in Corrêa et al.'s (2023) 2019 drill-down where Freedom / Autonomy / Democratic Values / Technical Sovereignty trumped Privacy. Additionally, despite the global view, there are regional variations. Asia places Beneficence / Non-maleficence fifth (74%), and Accountability / Liability sixth (70%). Better representation for China could change global values. Furthermore, Fjeld et al. (2020) identify theme bias based on western documents, where non-Western principles are considered outliers, like Japan's principle on fair AI competition.

Corrêa et al.'s paper ends with 2022, so does not reflect the new guidelines and regulations following generative AI. While the authors create mutually exclusive categories of government regulation, self-regulation and recommendation, as well as legally binding or non-binding, they do not represent the weight or impact of each document.

For instance, underscoring safety, bias and privacy principles, the Bletchley Declaration at the AI Safety Summit in November 2023 was the first international agreement across 28 countries, including the EU, US, and China, on addressing the opportunities and risks of frontier AI (GOV.UK, 2023).

EVALUATION

However, agreeing ethical principles at a global level still does not mean they meet local needs, interoperate, are applied, or are enforced. Practical AI ethics also requires monitoring and evaluating guideline impacts (Deckard, 2023).

Hagendorff (2020) describes self-governance as a tactic to appear compliant, while remaining vague to stave off protests, advance business interests, and avoid legally binding government regulations. Hagendorff and Corrêa et al. observe that significant male bias (roughly two-thirds) skews ethics guidelines toward existing technical solutions, rather than human-centric with a social or environmental focus.

I propose three recommendations:

- Regulation: Enforceable interoperable, global human-centric regulation.
- Social: Informed consent with fair compensation and attribution to creatives.
- Professional: Diversity and inclusion quotas for AI ethics boards.

The European Union (EU) AI Act, to be enacted in 2024, paves the way for enforceable human-centric risk-based regulation (European Parliament, 2024). Similar to GDPR, it is, so far, the only regulation that is human-centric, rather than innovation-centric (UK, USA, India) or state-focused (China) (Holistic AI, 2024). However, global regulation requires interoperability, and the rest of the world has voluntary standards. The OECD.AI policy observatory is one of the core places to track standards, tools, and incidents, however, navigating regulations is complex and legal expertise is helpful (OECD.AI, 2024).

In my industry, media, the Writers Guild of America (WGA) and SAG-AFTRA strikes focused on AI—including ensuring written material generated by AI is not classed as

human literary writing, and “informed consent” and “fair compensation” for digital replication (Cavna, 2023; Timsit, 2023). Maintaining human creativity is essential to long-term business strategies to avoid AI poisoning and model collapse due to continual training on AI-generated data (Rao, 2023).

Professionally, external AI ethics boards can provide diverse expert advisory with enforceable power. Unfortunately, Google fired two internal ethics researchers for raising lack of diversity concerns, whereas Meta’s external and independent Oversight Board includes geographical diversity bylaws (Schuett et al., 2024). Diverse independent oversight with legally enforceable obligations helps companies act ethically while minimising bias, but requires time, expertise, and commitment.

CONCLUSION

The generative AI inflection point has made ethical AI urgent for creatives and humanity. While global consensus is useful, practically, we need informed consent, fair compensation, attribution, enforcement, and interoperability, while meeting local needs and cultural differences. Recent advancements like the Bletchley Declaration and the EU AI Act are steps in this direction. However, there remains a balance between innovation and ethics, and there are practical challenges in enforcing, especially with voluntary standards. Independent and diverse representation is crucial, especially from underrepresented regions and groups. Ethical guidelines need to consider practical implementation and enforcement from a human-centric perspective. This requires an ongoing diverse dialogue among stakeholders in AI—policy, technology, public—to ensure guidance and regulation is created and implemented fairly for all.

REFERENCES

Carr, D.F. (2024) *ChatGPT Traffic up 13% YoY, Nearly Matching 2023 Peak*. Available from: https://www.similarweb.com/blog/insights/ai-news/chatgpt-rebuilds/ [Accessed 4 May 2024].

Cavna, M. (2023) *What actors won in the SAG-AFTRA strike deal*. Available from: https://www.washingtonpost.com/entertainment/2023/11/10/sag-strike-deal-details/ [Accessed 5 May 2024].

Corrêa, N.K., Galvão, C, Santos, J.W., Del Pino, C., Pinto, E.P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E. & de Oliveira, N. (2023) Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance, *Patterns* 4(10): 100857. DOI: https://doi.org/10.1016/J.PATTER.2023.100857.

Deckard, R. (2023) *What are ethics in AI? | BCS*. Available from: https://www.bcs.org/articles-opinion-and-research/what-are-ethics-in-ai/ [Accessed 5 May 2024].

European Parliament (2024) *Artificial Intelligence Act: MEPs adopt landmark law*. Available from: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law [Accessed 5 May 2024].

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. (2020) *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Available from: https://ssrn.com/abstract=3518482.

GOV.UK (2023) *Countries agree to safe and responsible development of frontier AI in landmark Bletchley Declaration*. Available from: https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration [Accessed 5 May 2024].

Hagendorff, T. (2020) The Ethics of AI Ethics: An Evaluation of Guidelines, *Minds and Machines* 30(1): 99–120. DOI: https://doi.org/10.1007/S11023-020-09517-8/TABLES/1.

Holistic AI (2024) *The State of Global AI Regulations in 2024*. Available from: https://www.holisticai.com/papers/the-state-of-ai-regulations-in-2024 [Accessed 5 May 2024].

Jobin, A., Ienca, M. & Vayena, E. (2019) The global landscape of AI ethics guidelines, *Nature Machine Intelligence 2019 1:9* 1(9): 389–399. DOI: https://doi.org/10.1038/s42256-019-0088-2.

OECD.AI (2024) *OECD AI Principles overview*. Available from: https://oecd.ai/en/ai-principles [Accessed 5 May 2024].

Rao, R. (2023) *AI-Generated Data Can Poison Future AI Models*. Available from: https://www.scientificamerican.com/article/ai-generated-data-can-poison-future-ai-models/ [Accessed 5 May 2024].

Russell, S. & Norvig, P. (2021) *Artificial Intelligence: A Modern Approach, Global Edition*. 4th ed. Pearson Education, Limited.

Schuett, J., Reuel, A.-K. & Carlier, A. (2024) How to design an AI ethics board, *AI and Ethics 2024* 1–19. DOI: https://doi.org/10.1007/S43681-023-00409-Y.

Timsit, A. (2023) *Hollywood studios and writers have a strike-ending deal. What's in it?* Available from: https://www.washingtonpost.com/style/2023/09/27/wga-contract-details-writers-strike-deal/ [Accessed 5 May 2024].