Fairness Without Reflection? An Empirical Analysis of AI Ethics Tools in the Age of GenAI

LISA KOERITZ, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Germany
SONJA PFISTERER, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Germany
MONE SPINDLER, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Germany
SIMON HIRSBRUNNER, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Germany

AI ethics tools designed to operationalise fairness are increasingly insufficient for the challenges posed by generative AI (GenAI). Based on a content analysis of 30 digital AI ethics tools, we find that tools function primarily as compliance artifacts rather than as epistemic devices supporting ethical judgment under change. Technical tools frame fairness as a computational problem inapplicable to GenAI, while governance-oriented tools reduce it to documentation and risk management. We argue that fairness in GenAI must be governed, not computed. This requires a shift toward tools that are adaptable to changing technical landscapes, clarify responsibilities across stakeholders, and support collaborative ethical reflection rather than static verification. This approach better serves the EU AI Act's requirements for human oversight than simple checklists, and ensures that fairness remains a continuous governance practice.

Keywords: algorithmic fairness, AI ethics tools, generative AI, AI governance, responsible AI, EU AI Act

Reference Format:
Lisa Koeritz, Sonja Pfisterer, Mone Spindler, and Simon Hirsbrunner. 2026. Fairness Without Reflection? An Empirical Analysis of AI Ethics Tools in the Age of GenAI. In *Proceedings of Fifth European Conference on Algorithmic Fairness (ECAF'26).* Proceedings of Machine Learning Research, 6 pages.

1 Introduction

Fairness has become a central concern in AI governance, particularly within the European regulatory landscape. In response, a wide range of AI ethics tools has emerged, promising to support responsible AI development through audits, assessments, metrics, and procedural guidance. The understanding of "tools" for operationalising AI ethics

Authors' Contact Information: Lisa Koeritz, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Tübingen, Germany, lisa.koeritz@uni-tuebingen.de; Sonja Pfisterer, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Tübingen, Germany, sonja.pfisterer@uni-tuebingen.de; Mone Spindler, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Tübingen, Germany, mone.spindler@uni-tuebingen.de; Simon Hirsbrunner, International Center for Ethics in the Sciences and Humanities, University of Tübingen, Tübingen, Germany, simon.hirsbrunner@uni-tuebingen.de.

This paper is published under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC-BY-NC-ND 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

ECAF'26, September 02–September 04, 2026, Ghent, BE

© 2026 Copyright held by the owner/author(s).

is broad, commonly defined as resources facilitating the translation of ethical principles into actionable practices [2, 9]. These include guidelines, checklists, audits, and software aimed at ensuring compliance. However, the rapid evolution of Generative AI (GenAI) fundamentally challenges these traditional fairness approaches. Static metrics and linear lifecycle checks, designed for supervised machine learning, are increasingly insufficient for black-box, emergent systems where risks are dynamic and context-dependent. While fairness remains the dominant principle in AI ethics operationalisation [12], growing body of critical literature argues that such operationalisations risk reducing fairness to a technical property of models, thereby obscuring its social, institutional, and normative dimensions and that focusing on values like fairness alone is insufficient if embeddedness into larger socio-technical contexts is ignored [11, 14]. Current tools often function as compliance artifacts that stabilise uncertainty for organisations, rather than as epistemic devices that support ethical analysis, judgment and operationalisation under change. This paper argues that fairness in GenAI cannot be "computed" alone; it must be governed. We therefore adopt an understanding of tools functioning as didactic arrangements, which share the aim of creating space, motivation and guidance for ethical reflection in innovation processes. We posit that holistic ethical reflection is not a replacement for fairness metrics, but a necessary instrument to achieve it in this new landscape.

2 Methodology

Within an interdisciplinary research project focusing on tools as means of ethics integration in the AI development process, we analysed a sample of 30 digital AI ethics tools. The full list can be found in Table 1. As part of a broader analysis of AI ethics tooling, we collected digital tools that aim to support the operationalisation of AI ethics and conducted a content analysis based on their approach to integrating ethics in the context of GenAI.

We based our initial collection on the substantial body of prior survey work regarding those tools starting with the extensive collection by [10] and expanding and referencing it with further collections (i.e., [1, 3, 5, 8, 10, 13–15]). To extend the temporal coverage of our tool collection, we conducted additional targeted screenings of recent conference proceedings (CHI, CSCW and FAccT 2024 and 2025) using keyword search and incorporated a small number of tools identified through expert recommendation.

As our research focused on readily available digital methods specifically, we applied a multi-criteria filtering. Only those tools were included in our sample that are:

- Digitally mediated (e.g. software, runnable package, web app, interactive workbook, or hosted interface), which we see as a requirement for the other criteria,
- Interactive, meaning offering a) a structured user interaction or executable processes, which results b) in some sort of tailored output generated, exported, compiled, logged, or instantiated by the tool (e.g., metrics, reports, completed cards, checklists, assessment outputs), which we see as a requirement for bringing ethical reflection to action, and
- Accessible, meaning direct interaction with the tool was possible without requiring paid access or restricted organizational login, which was on the one hand a requirement for our research but on the other hand also a precondition for equal access to the tool.

Within our overarching analysis, we analysed each tool regarding the scope of ethical engagement enabled in the context of GenAI systems. Across our collection of tools, however, we observed a distinct focus on fairness as a central aim within digital AI ethics tools.

Within our broader analysis, as a principle case study we thus analysed

- the scope of applicability in the GenAI context and
- the approach to fairness specifically with regard to how fairness is (explicity or implicitly) operationalized.

Especially the tools using technical metrics we analysed more closely used fairness as their primary lens for supporting AI ethics.

3 Key Findings

Drawing on our content analysis, we find a bifurcation in the tool landscape. First, technical tools largely frame fairness as a technical optimisation problem. These tools are predominantly designed for supervised machine learning and are not applicable to GenAI architectures where outputs are probabilistic and can exhibit emergent behaviour. Among them, widely used libraries such as IBM AI Fairness 360, Fairlearn, and TensorFlow Fairness Indicators do not even extend to GenAI approaches. In total, 20 of the 30 tools analysed are not applicable to the specific challenges of GenAI contexts at all. Second, documentation and assessment tools emerging for the GenAI landscape (such as the HEAT Toolkit, RAI Toolkit and Nokia AI Design Toolkit) use fairness, if at all, as a practically operationalisable lens, focusing on a risk management perspective. While notable examples such as the HEAT Toolkit attempt to incorporate a broader social justice context in their questioning, methodologically they still focus on self-assessment and documentation. Across both tool types, a shared pattern emerges: tools function primarily as compliance artifacts that stabilise organisational uncertainty, rather than as epistemic devices supporting ethical analysis, judgment and operationalisation under change. They, e.g., assume that the information and transparency required for fairness metrics exist, which is rarely the case in black-box GenAI systems, or centralized oversight and unified governance, which does not reflect the fragmented, multi-actor reality of modern GenAI infrastructures operating as distributed, nested software stacks, meaning that developers and deployers rarely have full control over all components. Principle-based operationalisation in digital form thus insufficiently captures the complexity of fairness in socio-technical systems: tools tend to reduce fairness to closing ethical questions (“Have we tested model results for fairness [. . .]”?, Deon Ethics Checklist; “Are your algorithms prejudiced?”, Digital Ethics Compass; “Did you consider diversity [. . .]?”, ALTAI-Tool) rather than structuring ongoing ethical reflection on context and social impact.

4 Discussion: Towards Holistic Fairness Governance

For stabilised, trustworthy technology development, we argue for a more holistic ethical approach. The different areas of ethics input need to be prioritised over rigid lifecycle discussions. We do not argue that lifecycle phases are obsolete, but that pure lifecycle-based checks are insufficient in a GenAI era. We need a shift from phase-based checks to continuous governance structures. Similarly, for fairness, tools need to enable substantial discussions instead of merely producing or analysing metrics. This demands a rethinking of tool design along two dimensions. First, tools need to be adaptable and editable, capable of responding to changing technical landscapes rather than encoding static assumptions about system behaviour. Second, tools need to support collaborative ethical reflection: clarifying responsibilities across stakeholders, fostering ethical reflection competence, and aiming for a holistic understanding of socio-technical impact rather than individual compliance checkpoints. This approach serves the EU AI Act’s requirements for human oversight and risk management more in depth than simple checklists [6]. The

Table 1. Fairness operationalization across AI ethics tools

Tool	Fairness approach	Applicable to GenAI
Unbias AI (Aequitas)	Fairness metrics	
AI Explainability 360	No fairness approach	
Fairlearn	Fairness metrics	
InterpretML	No fairness approach	
Learning Interpretability Tool (LIT)	Measurable objectives	✓
NLP Checklist	Bias probing	
TensorFlow Fairness Indicators	Fairness metrics	
What-if Tool	Fairness metrics	✓
AI Fairness 360	Fairness metrics	
Model Cards Toolkit	Documentation	✓
Data Cards Playbook	Documentation	
High-risk EU AIAct Toolkit	Self-assessment	✓
ELSI-SAT Health and Care	Guided reflection	
Cards for Humanity	Guided reflection	
The Digital Ethics Compass	Guided reflection	
Data Ethics Emergency Drill	No fairness approach	✓
Farsight	No fairness approach	✓
AI Ethics Quiz	Guided reflection	
The Box	Self-assessment	
HAX Toolkit (Workbook)	Self-assessment	✓
Deon Ethics Checklist	Self-assessment	
AI Maturity Tool	No fairness approach	
ALTAI Tool	Self-assessment	
Nokia AI Design Toolkit	Self-assessment	✓
RAI Toolkit	No fairness approach	✓
Responsible AI Diagnostic	Self-assessment	
AI System Ethics Self-Assessment	No fairness approach	
Mission KI Pruefportal	Self-assessment	✓
Algorithmic Impact Assessment	Self-assessment	
AI Readiness Radar	Self-assessment	

AI Act mandates ongoing monitoring for high-risk systems; without tools enabling substantive reflection, such monitoring risks remaining superficial [7]. Auditing is therefore necessary, but not sufficient. Holistic reflection is needed to address systemic societal impacts that metrics miss. Only through tools enabling deep ethical reflection can fairness in GenAI be dynamically secured.

This paper contributes a critical perspective on the current design and function of AI ethics tools in Europe. Current AI ethics tools largely fail to enable substantive ethical reflection for fairness in the context of GenAI. As a principle, fairness remains a highly debated [4], yet apparently highly actionable concept. We call on researchers and practitioners to prioritise tools that enable reflection over tools that promise compliance. Fairness must be understood as a continuous governance task, supported by tools that facilitate collaborative ethical reflection rather than static verification. We envision, this would allow to not only ask and narrow conversation regarding fairness toward how to mitigate based on predetermined notions but allow space to open up and guide an embedded idea of context-based negotiation of the normative idea of fairness. By shifting focus from technical metrics to holistic governance, we can better address the systemic risks of GenAI and ensure that fairness remains a living practice rather than a procedural compliance exercise.

Acknowledgments

This research is part of the research and development project ANKER which was funded by the German Federal Ministry of Research, Technology and Space (BMFTR) within the funding number 16SV9223 and managed by the Project Management Agency VDI/VDE Innovation + Technik GmbH. The authors are responsible for the content of this publication.

References

[1] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* 2, 3 (Aug. 2022), 405–429. doi:10.1007/s43681-021-00084-x

[2] Vita Santa Barletta, Danilo Caivano, Domenico Gigante, and Azzurra Ragone. 2023. A Rapid Review of Responsible AI frameworks: How to guide the development of ethical AI. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*. ACM, Oulu Finland, 358–367. doi:10.1145/3593434.3593478

[3] Glen Berman, Nitesh Goyal, and Michael Madaio. 2024. A Scoping Study of Evaluation Practices for Responsible AI Tools: Steps Towards Effectiveness Evaluations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–24. doi:10.1145/3613904.3642398

[4] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *Comput. Surveys* 56, 7 (July 2024), 1–38. doi:10.1145/3616865

[5] Keeley Crockett, Edwin Colyer, Luciano Gerber, and Annabel Latham. 2023. Building Trustworthy AI Solutions: A Case for Practical Solutions for Small Businesses. *IEEE Transactions on Artificial Intelligence* 4, 4 (Aug. 2023), 778–791. doi:10.1109/TAI.2021.3137091

[6] European Parliament. 2024. AI Act Regulation (EU) 2024/1689 :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 (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). https://data.europa.eu/doi/10.2804/2823574

[7] Jan Gogoll and Niina Zuber. 2026. *Introduction to Ethical Software Development*. Springer Nature Switzerland, Cham. doi:10.1007/978-3-032-06195-9

[8] Blaine Kuehnert, Rachel Kim, Jodi Forlizzi, and Hoda Heidari. 2025. The “Who", “What", and “How" of Responsible AI Governance: A Systematic Review and Meta-Analysis of (Actor, Stage)-Specific Tools. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens Greece, 2991–3005. doi:10.1145/3715275.3732191

[9] Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal. 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 (Aug. 2020), 2141–2168. doi:10.1007/s11948-019-00165-5

[10] Ricardo Ortega-Bolaños, Joshua Bernal-Salcedo, Mariana Germán Ortiz, Julian Galeano Sarmiento, Gonzalo A. Ruz, and Reinel Tabares-Soto. 2024. Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence*

Review 57, 5 (April 2024), 110. doi:10.1007/s10462-024-10740-3

[11] Tijs Portegies, Oumaima Hajri, Lotte Willemsen, and Maaike Harbers. 2025. Decoding AI ethics: What do ethical tools tell us about ethics? *Ethics and Information Technology* 27, 4 (Dec. 2025), 65. doi:10.1007/s10676-025-09860-3

[12] Shalaleh Rismani, Leah Davis, Bonam Mingole, Negar Rostamzadeh, Renee Shelby, and AJung Moon. 2025. Responsible AI measures dataset for ethics evaluation of AI systems. *Scientific Data* 12, 1 (Dec. 2025), 1980. doi:10.1038/s41597-025-06021-5

[13] Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. 2024. Guidelines for Integrating Value Sensitive Design in Responsible AI Toolkits. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20. doi:10.1145/3613904.3642810

[14] Richmond Y. Wong, Michael A. Madaio, and Nick Merrill. 2023. Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–27. doi:10.1145/3579621

[15] Jianlong Zhou and Fang Chen. 2026. AI Ethics Operationalisation: Progress, Tools, and Opportunities. doi:10.2139/ssrn.6067748