



Fairness Without Reflection?

An Empirical Analysis of AI Ethics Tools in the Age of GenAI

Lisa Koeritz, Sonja Pfisterer, Dr. Mone Spindler, Dr. Simon Hirsbrunner

International Center for Ethics in the Sciences and Humanities (IZEW), University of Tübingen



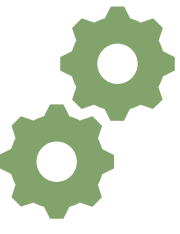
Background

Research Project „ANKER“

- Research project focusing on tools as means of ethics integration in the AI development process.
- Result:** Analysis of AI ethics tools as well as prototypical development and evaluation of new approaches to AI ethics tooling.

Digital AI Ethics Tools <> Fairness

- Our Focus: **Tools** functioning as didactic arrangements, which share the aim of creating space, motivation and guidance for ethical reflection in innovation processes.
- Fairness** remains the dominant principle in AI ethics operationalisation.

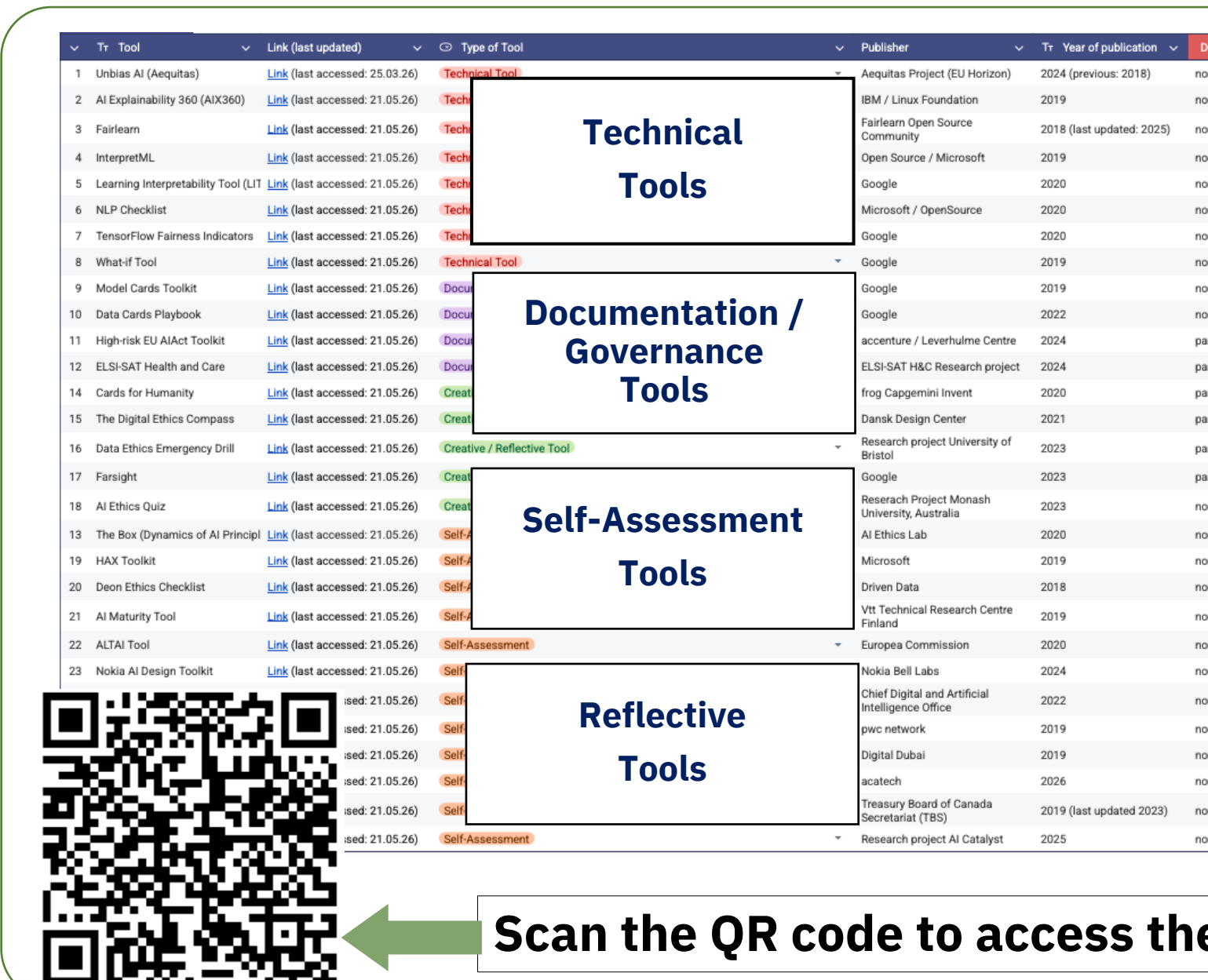


Methodology

Digital AI Ethics Tools

Tools that aim to support the operationalisation of AI ethics

- ✓ **Digitally mediated:** digital interface with output
- ✓ **Interactive:**
 - a structured user interaction or executable processes, which results
 - in some sort of tailored output generated, exported, compiled, logged, or instantiated by the tool
- ✓ **Accessible:** not requiring paid access or restricted organizational login



Content Analysis of 30 Digital AI Ethics Tools

- Collection of digital tools that aim to support the operationalisation of AI ethics
- Content analysis based on tools approach to anchor ethics in the AI development process and evaluation of each tool along different dimensions

GenAI Applicability

Examined the extent to which tools address challenges specific to generative AI. Coding focused on whether tools explicitly mention or engage with GenAI-specific considerations in their design, framing, or guidance.

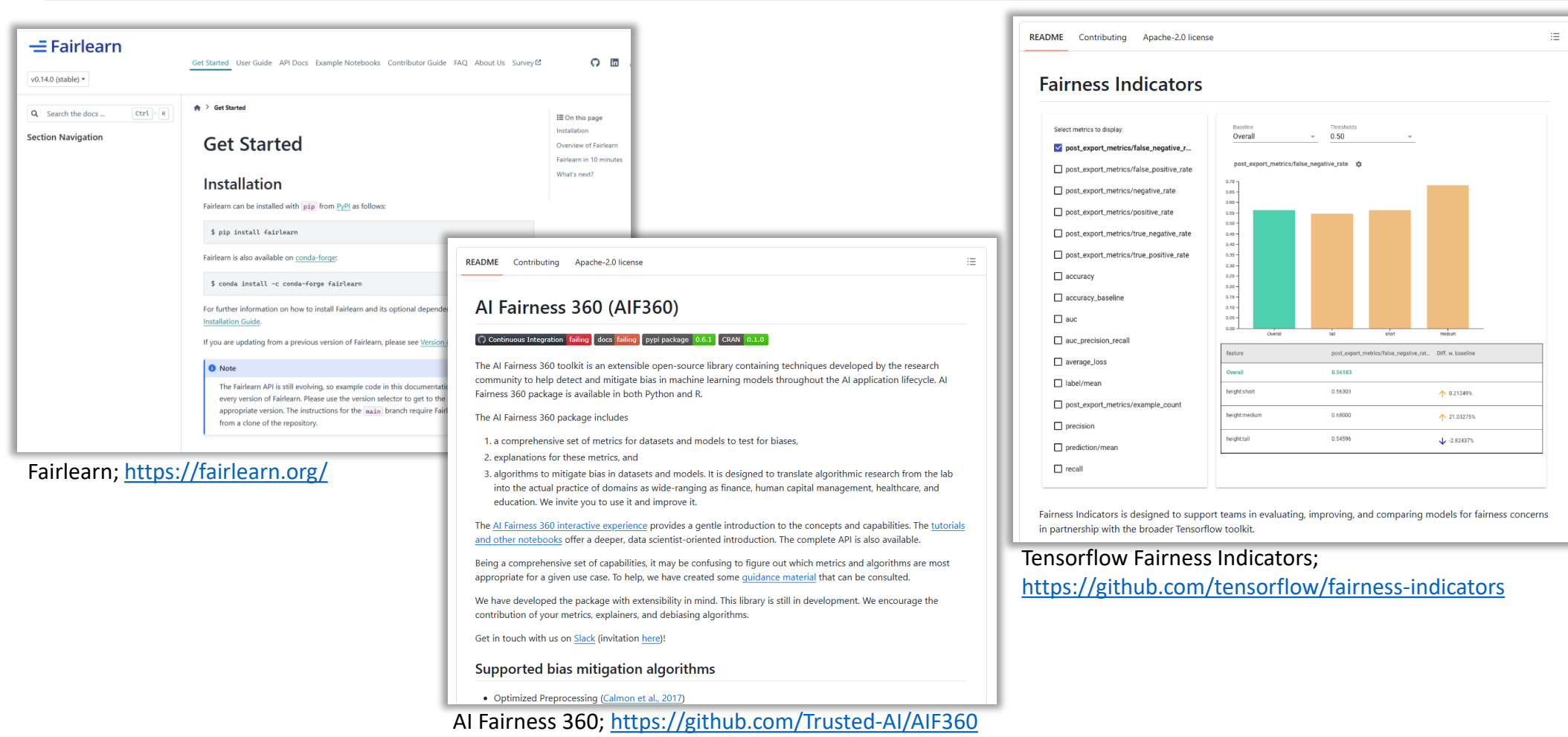
Fairness Operationalization Approach

Analysed whether fairness was specifically assessed, documented, probed or measured for or reflected on within the tool.



Results

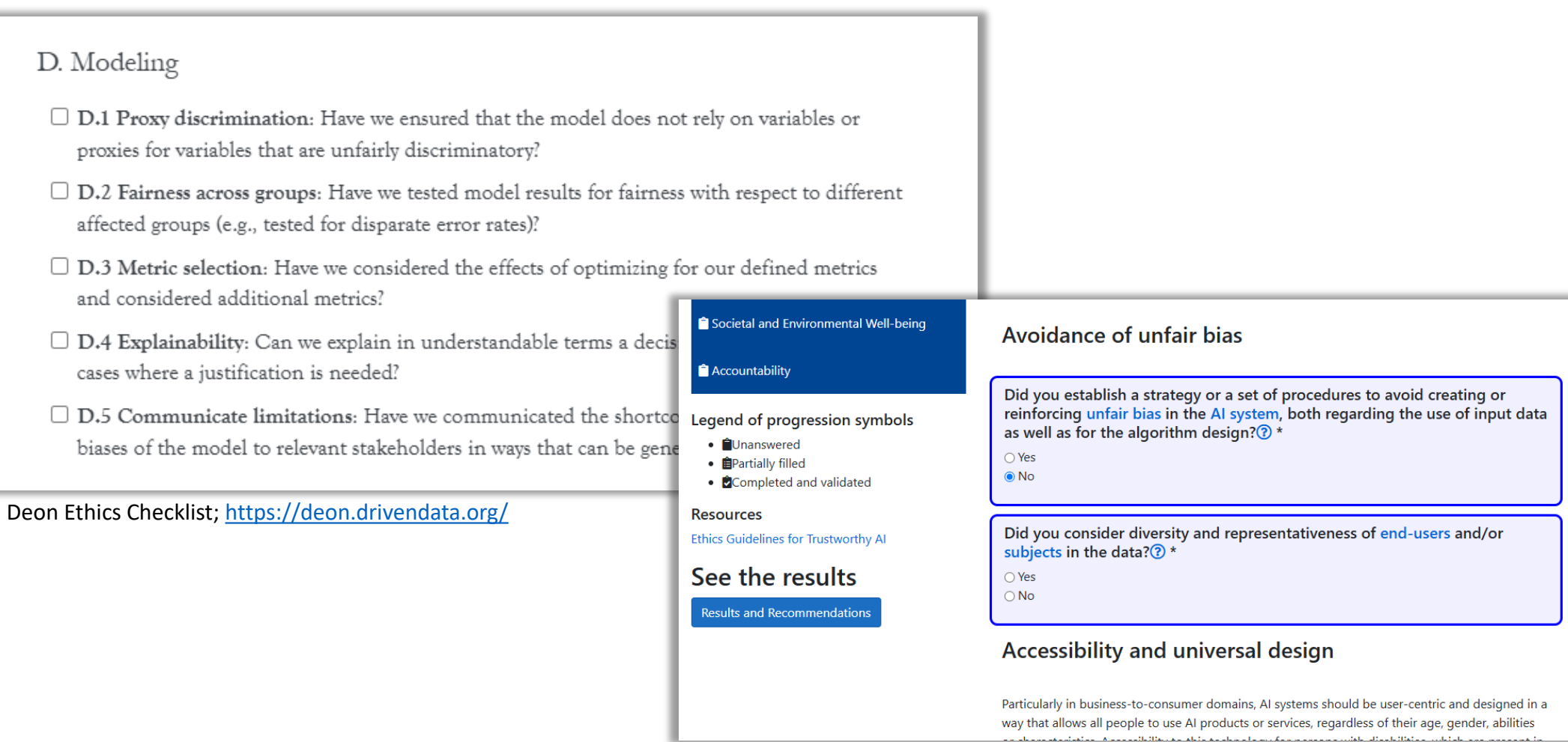
Fairness As a Technical Optimisation Problem



Tools tend to assume that the *information and transparency required for fairness metrics* or a *centralized oversight and unified governance* exists.

Challenge: Fragmented multi-actor reality in modern GenAI infrastructures where developers and deployers rarely have full control over all components.

Fairness as a Risk Management Problem



Tools tend to *reduce fairness to closing ethical questions*.

Challenge: Questions do not structure ongoing ethical reflection on context and social impact.

Tools function primarily as compliance artifacts that stabilise organisational uncertainty, rather than as epistemic devices supporting ethical analysis, judgment and operationalisation under change.

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 AI Act 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	

Discussion: Toward Holistic Fairness Governance

- Pure lifecycle-based checks are insufficient in a GenAI era: Shift from phase-based checks to continuous governance structures is needed
- AI Act mandates ongoing monitoring for high-risk systems: Without tools enabling substantive reflection, such monitoring risks remaining superficial
- Auditing is therefore necessary, but not sufficient: Holistic reflection is needed to address systemic societal impacts that metrics miss.

Rethinking of Tool Design

- Tools need to be adaptable and editable:** Tools should allow continuous adaptation toward changing technical landscape rather than static assumption
- Tools need to support collaborative ethical reflection:** Tools should clarify responsibilities across stakeholders, foster ethical reflection competence, and aim for a holistic understanding of the socio-technical impact

Conclusion

Fairness must be understood as a continuous governance task, supported by tools that facilitate collaborative ethical reflection rather than static verification.

By allowing space to open up and guide an embedded idea of context-based negotiation of the normative idea of fairness, we can better address the systemic risks of GenAI and ensure that fairness remains a living practice rather than a procedural compliance exercise.

References (Selection):

Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elthalal. 2020. From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Train Principles into Practices. *Science and Engineering Ethics* 26, 4 (Aug. 2020), 2141–2168. doi:10.1007/s11948-019-00165-5

Shalaleh Rismeni, Leah Davis, Bonnam 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), 1990. doi:10.1038/s41597-025-06021-5

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

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

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

Hagedorn, T. 2024. Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *Minds and Machines*, 34(4).

Raman, V.; Shen, J. H.; Zhang, A. K.; Gailmard, L.; Bomasani, R.; Ho, D. E.; and Wang, A. 2025. Disclosure and Evaluation as Fairness Interventions for General-Purpose AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(3): 2121–2135

Gefördert durch:



Bundesministerium
für Forschung, Technologie
und Raumfahrt

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.

<https://uni-tuebingen.de/de/257900>

