Edukacja Filozoficzna 79/2025
ISSN 0860-3839
DOI: 10.14394/edufil.2025.0001
ORCID: 0000-0002-5506-9027
ORCID: 0000-0002-6336-1036
ORCID: 0000-0001-7159-3497

AI Ethics beyond Compliance: Governance, Power, and Human Flourishing

Sara Lumbreras
(Instituto de Investigación Tecnológica, Universidad Pontificia Comillas)

Andrea Vestrucci
(Department of Computer Science, Universität Bamberg;
Christ School of Theology)

Ralph Stefan Weir
(School of Humanities and Heritage, University of Lincoln)

Artificial intelligence (AI) is rapidly being integrated across society and is increasingly used in a wide spectrum of decision-making processes, from business operations to public service allocation, healthcare support, credit scoring, and recruiting. In particular, large language models (LLMs) have become commonplace in educational institutions and workplaces, and are increasingly influencing everyday communication practices, including their use as companions or supports for loneliness.[1]

In light of AI's growing presence in our lives, there has been a notable rise in documents and publications deepening the ethical aspect of AI, ranging from organizational policies and corporate guidelines to global initiatives. Here we focus on three examples. In 2021, UNESCO adopted the non-binding *Recommendation on the Ethics of Artificial Intelligence*, which lays out principles and

[1] A. de Wynter, *If Eleanor Rigby Had Met ChatGPT: A Study on Loneliness in a Post-LLM World*, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, eds. W. Che et al., Vol. 1, Association for Computational Linguistics, Vienna 2025, pp. 19898–19913.

calls on member states to implement policy measures across the AI lifecycle.[2] In 2024, the European Union officially adopted the AI Act, establishing the first comprehensive legal framework for AI and introducing a risk-based classification of AI systems.[3] In 2025, the Australian government updated its policy for the responsible use of AI, which sets requirements for how Australian government agencies should adopt and govern AI.[4]

This growing attention to ethics is encouraging, but it also risks reducing ethical engagement to mere legal or procedural compliance. There is a persistent concern about "ethics washing,"[5] whereby institutions and companies deploy ethical language to maintain their reputations without making substantial changes in practice. In such settings, operational questions about what is good, just, or fair grounded in lived human experience tend to be neglected. Moreover, although issues of fairness, well-being, ecological sustainability, privacy, and inclusion are widely recognized as core concerns, they are often treated in fragmented ways and bundled under broad labels and "buzzwords" like "trustworthiness" or "responsibility."[6]

This special issue brings together perspectives from across disciplines and traditions to explore how AI ethics is shaped by governance frameworks, societal institutions, educational practices, and contested ideas of justice and agency.

The relationship between ideology and power is critically examined in Luka Perušić's article, *Ideological Limits to Ethical Artificial Intelligence*. Perušić explores how the concept of "ethical AI" is shaped, and often constrained, by underlying ideological commitments. He argues that despite the proliferation of ethical guidelines and value-alignment frameworks, the ethical often functions as a malleable label within corporate, regulatory, and geopolitical contexts,

[2] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, SHS/BIO/REC-AIETHICS/2021, URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

[3] European Commission, *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)*, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689.

[4] Australian Government, *Policy for the Responsible Use of AI in Government*, version 2.0, URL: https://www.digital.gov.au/ai/ai-in-government-policy.

[5] See G. van Maanen, *AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics*, "Digital Society" 2022, Vol. 1, 9, https://doi.org/10.1007/s44206-022-00013-3.

[6] See Karoline Reinhardt's comprehensive critique of the term "trustworthiness" in the field of AI ethics in K. Reinhardt, *Trust and Trustworthiness in AI Ethics*, "AI and Ethics" 2023, Vol. 3, pp. 735–744, https://doi.org/10.1007/s43681-022-00200-5.

vulnerable to "ethics washing" and competing social preferences. By analyzing the status of ethical claims in current governance instruments, the paper shows how ideological structures set practical limits on what ethical AI can achieve, and how these limits must be acknowledged in any realistic theory of responsible AI.

Education is revisited in *Computational Analysis for Philosophical Education: A Case Study in AI Ethics*, which applies natural language processing to analyze AI ethics syllabi. Alex Cline, Brian Ball, David Peter Wallis Freeborn, Alice C. Helliwell, and Kevin Loi-Heng investigate what contemporary natural-language-processing techniques can reveal about the content and structure of AI ethics curricula. They demonstrate how computational methods can bring conceptual patterns to the surface, highlight thematic emphases, and support pedagogical reflection. The paper situates this approach within the digital humanities and proposes computational analysis as a promising resource for philosophical teaching and curriculum design.

Neomal Silva's contribution, *Justice and AI Fairness: John Rawls and Iris Marion Young on Racist and Sexist AI Decisions*, centres justice as a response to structural oppression. Drawing on cases of algorithmic bias (such as discriminatory hiring tools and flawed facial recognition) Silva critiques the limitations of Rawlsian distributive justice and instead turns to Young's model of structural injustice. We cannot be content knowing that the "average" result is good for an algorithm, if a group is disproportionately damaged by its application. As an alternative, the paper turns to Young's critical theory, which incorporates structural power and consciousness-raising practices, arguing that her approach better captures the mechanisms through which discriminatory patterns are reproduced in machine-learning systems.

The theme of care, responsibility, and human–AI cooperation is explored further in *A Philosophical Account of Shared Autonomy and Moral Agency in Human–AI Teams*. Max Parks examines how agency becomes distributed across humans and machines in contexts ranging from autonomous vehicles to social robots. Parks argues that computational optimization cannot substitute for the socially embedded moral understanding characteristic of human judgement, and advances a care-theoretic framework for evaluating hybrid systems, emphasizing attentiveness, dependency, and relational accountability. Through cases such as self-driving vehicle scenarios and companion-robot interactions, the paper proposes principles for integrating AI in ways that enhance, rather than erode, meaningful human agency.

The special issue also interrogates how AI shapes the politics of knowledge. In the paper *In Defence of LLM-Based Tools in Scientific Writing: Epistemic and Ethical Considerations of LLM-Restrictive Publishing Policies*, Aleksandra Vučković analyzes the emerging tendency among universities and publishers to prohibit or severely limit the use of LLMs in academic writing. Vučković argues that current detection tools produce both false positives and false negatives, raising serious epistemic and professional risks, especially for non-native English-speaking researchers, who face disproportionate rates of mistaken suspicion. The article proposes a more moderate regulatory approach that recognizes both the linguistic benefits LLMs can provide and the limits of existing detection technologies.

A significant contribution arises from the dialogue between religious and secular approaches to AI governance. In *Ethical Evaluation of Artificial Intelligence from the Perspective of the Catholic Church*, Krzysztof Trębski analyzes the Catholic ethical evaluation of AI and the risks of unregulated development through documents of the Holy See, and the teaching and public pronouncements of recent pontiffs. Drawing on papal encyclicals, Vatican documents, and global policy instruments, the paper explores how AI development serves the dignity of the human person and the universal common good by tracing points of convergence and divergence between secular and ecclesial frameworks, particularly around autonomy, beneficence, and justice.

Taken together, these articles treat AI ethics not as an abstract list of principles, but as a domain rooted in social structures, interpersonal relationships, and power dynamics. They raise critical, practical questions: Who benefits – and who bears the costs – when AI systems are deployed? Whose perspectives inform design and implementation choices, and whose are excluded? How is responsibility and care distributed across human–machine interactions, and how do institutions influence AI's development and use? A central concern explored by the special issue is vulnerability, whether in the experience of communities affected by biased systems, groups underrepresented in global governance debates, or scholars exposed to inequalities through language and publication practices. This volume exemplifies a genuinely interdisciplinary dialogue. Bringing critical theories of justice into conversation with feminist care ethics, Catholic social teaching, epistemology, and computational methodologies, it shows what becomes visible when AI ethics is approached from multiple standpoints and diverse perspectives. The aim is not to settle these debates, but to invite ongoing reflection and collective action towards the common good in an AI-driven world. Much more work remains, and it will need to be interdisciplinary if AI ethics is to meaningfully shape the development of this technology in ways that foster human dignity and encourage human flourishing.

Bibliography

Australian Government, *Policy for the Responsible Use of AI in Government*, version 2.0, URL: https://www.digital.gov.au/ai/ai-in-government-policy.

De Wynter A., *If Eleanor Rigby Had Met ChatGPT: A Study on Loneliness in a Post-LLM World*, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, eds. W. Che et al., Vol. 1, Association for Computational Linguistics, Vienna 2025, pp. 19898–19913.

European Commission, *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)*, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689.

Maanen G. van, *AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics*, "Digital Society" 2022, Vol. 1, 9, https://doi.org/10.1007/s44206-022-00013-3.

Reinhardt K., *Trust and Trustworthiness in AI Ethics*, "AI and Ethics" 2023, Vol. 3, pp. 735–744, https://doi.org/10.1007/s43681-022-00200-5.

UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, SHS/BIO/REC-AIETHICS/2021, URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455.