Edukacja Filozoficzna

International Journal of Philosophical Education

79/2025

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

*Edited by Andrea Vestrucci, Sara Lumbreras, Ralph Stefan Weir*



Wydział Filozofii Uniwersytetu Warszawskiego

**Wydawca / Publisher**

Uniwersytet Warszawski, Wydział Filozofii



Wersja drukowana / Printed version: ISSN 0860-3839
Wersja elektroniczna / Electronic version: ISSN 2956-8269

**Nakład / Print run:** 110 egz./copies

**Adres redakcji / Address:** ul. Krakowskie Przedmieście 3, p. 302, 00-927 Warszawa
**E-mail:** edukacjafilozoficzna@uw.edu.pl

**Strona internetowa / Website:** edukacja-filozoficzna.uw.edu.pl

Wszystkie teksty oznaczone jako artykuły naukowe przeszły pozytywnie procedurę recenzyjną opisaną na stronie internetowej. / All papers marked as scholarly articles have undergone rigorous peer review, as described on the journal's website.

Czasopismo „Edukacja Filozoficzna" znajduje się w wykazie czasopism naukowych MNiSW (40 punktów). / *Edukacja Filozoficzna* features in the Ministry of Science and Higher Education's index of scholarly journals (40 points).

Publikacja dofinansowana przez Uniwersytet Warszawski w ramach działania „Poprawa zdolności publikacyjnych" realizowanego w ramach programu „Inicjatywa Doskonałości – Uczelnia Badawcza". Publication has been subsidized by the University of Warsaw under the "Improving publication capacity" action implemented as part of the programme "Excellence Initiative – Research University."

# Table of Contents

Edukacja Filozoficzna
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.

Edukacja Filozoficzna
ISSN 0860-3839
DOI: 10.14394/edufil.2025.0002
ORCID: 0000-0002-5339-781X

# Ideological Limits to Ethical Artificial Intelligence

Luka Perušić
(University of Zagreb, Faculty of Humanities and Social Sciences)

**Abstract:** The paper analyzes the current use of *ethical* artificial intelligence (AI), argues that there are ideological limits to it, and discusses these limits. The topic is of particular relevance to research on the social implementation of AI systems, as ideological underpinnings are not easy to identify and ideology research is underrepresented in research on AI phenomena. The first section analyzes what counts as ethical in ethical AI systems. The second section classifies the dimensions of the ethical in AI systems, highlights their interrelationships and applies *forness* as a key concept that helps narrow the focus on the ideological component of ethical AI. The third section describes the presence of ideology in ethical AI and clarifies the limits it imposes on AI as a general phenomenon that undoubtedly has the potential to contribute to a more humane society, but is severely constrained by ideology.

**Key words:** morality, ethic, ethics, politics, ideology, forness, artificial intelligence, artificial agency

## 1. Understanding the "Ethical" in Ethical Artificial Intelligence[1]

### 1.1. The Status of the Ethical

In recent years most of the economically leading countries, supranational entities, and international expert organizations, together with the most influential technology companies, are striving to create workable frameworks for the development, implementation, use and evaluation of artificial intelligence (AI) systems. The level at which AI has suddenly been taken seriously is technologically unmatched, with the European Union (EU) at the forefront in terms of intensity, scope and thoroughness, culminating in the proposal of the Artificial Intelligence

[1] I sincerely thank the reviewers and editors for their efforts to advance the quality of the paper.

Act.[2] In the spectacle of initiatives, guidelines, strategies, policies and legislative preparations to harness the advances in AI development and implementation, the question of ethics has been ever present, and the notion of ethical AI could neither be avoided nor evaded. Scholars working in the fields of morality, ethics and AI, as well as policy makers and jurists dealing with ethical AI, have approached these problems with different emphases:

> [Strategies for approaching (ethical) AI] range from regulations, law, codes of conduct, attempts to design AI with safety and ethics uppermost, attempts to build ethics into design process, specific strategies such as attempts to understand and mitigate bias, and so on. Some focus on current issues; some focus on longer-term and more speculative questions, such as possible dangers of superintelligence. Some issues are concrete and specific; some are more general, wide-ranging, or foundational. Some approaches lean towards the view that AI presents a threat that we might lose control of ourselves and of our values and that we need radical shifts to deal with the world that is coming. Other approaches are more sanguine and diligently tread the path of trying to ensure that the technologies that are being developed and used fit within current frameworks of value in approaches broadly labelled "value alignment."[3]

However, the discourse on being "ethical" continues to most commonly perpetuate the idea that *ethical*[4] refers to having a set of principles that instruct on proper conduct towards others or on what values should be embodied and manifested. Although the approach may seem functional, in the contemporary technological forefront society *the ethical* was never made fundamental neither in terms of nurturing and education nor in terms of legislation – in the context of the triple helix complex (military, industry, academia), it was systematically relegated to "playing the role of a bicycle brake on an international airplane"[5]

---

2 European Parliament, P9_TA(2023)0236: Artificial Intelligence Act, June 2023, amendments; cf. Council of the European Union, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts: Analysis of the Final Compromise Text with a View to Agreement*, no. Cion doc. 8115/21, Brussels, 26 January 2024.

3 P. Boddington, *AI Ethics: A Textbook*, Springer, Singapore 2023, p. 6.

4 Almost always derived from the word *ethics*, even though it should derive from *ethic*, as the latter is a set of action-guiding principles related to moral behaviour, while the former is a branch of philosophy.

5 U. Beck, *Gegengifte. Die organisierte Unverantwortlichkeit*, Suhrkamp Verlag, Frankfurt am Main 1988, p. 194.

already in the 1980s. Contemporary studies support this argument by showing that ethical norms have *near-zero* influence on the tech community, both the student population and working experts.[6] This raises the possibility that incentives from major players to build an ethical AI and use AI ethically may not be ethical and may not support the consistency between moral behaviour and legislation. A brief insight into the motives behind the formation of ethically aligned products was given in 2016 by a Mercedes representative, Christoph von Hugo, whose comment was one of the first public comments on the moral issues related to self-driving cars made by companies producing such vehicles. Von Hugo stated that self-driving Mercedes cars "would always prioritize their owners," before changing his statement after a public outcry.[7] It is a preference that is understandable from the perspective of a product seller, but not from the perspective of the fundamental rules and laws of traffic regulation or from the perspective of non-driving members of the contemporary social environment. The relegation of the ethical to an inferior position produced at least two consequences: (1) the possibility to manipulate the notion of the ethical for the protection of personal gain against the other, and (2) the relativization of morality.

In relation to the first consequence, the devaluation of the ethical has reached a new level, especially in the context of climate change and sustainability, with technology companies themselves expressing ethical concerns for twofold effect: (1) *legislative*, because by pretending to deal with the ethics of their own inventions they are stalling talks about the regulation of their activities and products,[8] and (2) *commercial*, because by expressing their ethical standpoints and pasting them on the "cover" of their brand they market themselves as trustworthy humanists contributing to the creation of a better society, and thus a better label for stakeholders to spend money on.[9] The epitome of this misleading strategy is their

---

6 L. Munn, *The Uselessness of AI Ethics*, "AI and Ethics" 2023, Vol. 3, No. 3, p. 872, https://doi.org/10.1007/s43681-022-00209-w; T. Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, "Minds and Machines" 2020, Vol. 30, No. 1, p. 108, https://doi.org/10.1007/s11023-020-09517-8. Both papers link to several different studies supporting the argument empirically.

7 S. Nyholm, *The Ethics of Crashes with Self-Driving Cars: A Roadmap, I*, "Philosophy Compass" 2018, Vol. 13, No. 7, e12507, p. 5, https://doi.org/10.1111/phc3.12507.

8 L. Munn, *The Uselessness of AI Ethics*, op. cit., p. 872.

9 N. de Marcellis-Warin et al., *Artificial Intelligence and Consumer Manipulations: From Consumer's Counter Algorithms to Firm's Self-Regulation Tools*, "AI and Ethics" 2022, Vol. 2, No. 2, p. 264, https://doi.org/10.1007/s43681-022-00149-5.

private, unregulated development of AI solutions, veiled by their publicly expressed concern about the apocalyptic coming of artificial general intelligence – a self-aware, adapting, and learning autonomous AI that may take human beings "out of the picture." The practice is as absurd as publicly warning that all of human civilization could die from a deadly virus while privately developing it in the lab; however, misleading and entertaining the public serve the purpose of allowing the companies the freedom to develop AI systems for these companies' gain.

In addition, Thilo Hagendorff highlights that "ethics can also simply serve the purpose of calming critical voices from the public, while simultaneously the criticized practices are maintained within the organization."[10] Deconstructed, the practice is a form of "ethics washing," a sibling to the well-known phenomenon of greenwashing. Ethics washing is "the practice of visibly, sometimes ostentatiously, showing to the world that one is taking great care to attend to ethics, while in reality, doing little or nothing."[11] A notorious example of such behaviour is the inappropriate firing of Timnit Gebru by Google after Gebru insisted on publishing a report that demonstrated how AI systems could generate racial results, while simultaneously presenting the company as a leader in ethical standards.[12] Granted, it would be unconvincing to claim that ethics washing – and the entirety of ethical devaluation processes – apply to the entirety of the AI systems development landscape. AI development spans from developing systems that assist in mountain rescue missions, through news feeds, to unmanned ground vehicles for assault combat, and there are certainly many authors who actively engage in discussions of the best possible utilization of AI systems. For example, Seng W. Loke, who, analyzing the game theory problem in the context of interaction among autonomous AI systems, has proposed the prime rule "Cooperate first" as "a good candidate for a universalizable maxim (i.e. 'seeking first to cooperate' could be willed as a strategy for everyone)" that would possibly manage the autonomous interaction of AI systems in a vehicle network for the benefit of all participants.[13] However, here I focus on examples generated by

[10] T. Hagendorff, *The Ethics of AI Ethics*, op. cit., p. 100.

[11] P. Boddington, *AI Ethics*, op. cit., p. 21; cf. L. Munn, *The Uselessness of AI Ethics*, op. cit., p. 872.

[12] For a good overview of the case and the understanding of multiple layers of misconduct involved in the process of firing Gebru, see T. Simonite, *What Really Happened When Google Ousted Timnit Gebru*, Wired, 8.06.2021, URL: https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/.

[13] S.W. Loke, *Designed to Cooperate: A Kant-Inspired Ethic of Machine-to-Machine Cooperation*, "AI and Ethics" 2022, Vol. 3, No. 3, p. 992, https://doi.org/10.1007/s43681-022-00238-5.

powerful global entities that might not be as clear as they initially appear. For example, although Microsoft as a global tech company is one of the parties that have committed to apply UNESCO's *Recommendation on the Ethics of Artificial Intelligence* in 2022,[14] it has invested over $13 billion into OpenAI, which has already been shown to favour exploitative practices.[15]

A second consequence of the relegation of the ethical to an inferior position is the multiplication of various ethical codes that are constantly proposed regardless of other efforts, resulting in a plethora of ethical proposals that are not supported by legal systems in terms of sanctions. Combined with the evidence of cultural differentiation in the world, a relativistic image of ethics emerges and a negative view of ethics as arbitrary or limited. Given the case, some authors argue that it is pointless to discuss the ethics of AI (and thus ethical AI) and that we should focus on law-abidingness and accountability.[16] Roman V. Yampolskiy raises the old but standing problem of *legal positivism* or *legal blindness*, in the sense that what is allowed or forbidden by law may be unethical (for instance, ban on same-sex marriage and acceptance of underage marriage), and later cannot be prosecuted because it was acceptable from the perspective of the law that was in effect at the time. In that regard, the EU's Artificial Intelligence Act is a peculiar case which serves well to clarify what "ethical" stands for in the phrase "ethical AI."

## 1.2. The Meaning of the Phrase "Ethical AI"

The European Commission (EC) has accepted a document titled *Ethics Guidelines for Trustworthy AI*, which considers "ethical AI" to be an AI system following the set of principles labelled by the Commission as "ethical" (respect for human autonomy, prevention of harm, fairness, and explicability).[17] This selection

[14] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris 2022.

[15] More on this in the next section. For the information on Microsoft, see J. Liboreiro, *European Regulators Put Microsoft's $13 Billion Bet on OpenAI under Closer Scrutiny*, EuroNews, 9.01.2024, URL: https://www.euronews.com/my-europe/2024/01/09/european-regulators-put-microsofts-13-billion-bet-on-openai-under-closer-scrutiny.

[16] For example, computer engineer Roman V. Yampolskiy, who stated this before the political world took AI seriously. See R.V. Yampolskiy, *Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach*, in: *Philosophy and Theory of Artificial Intelligence*, ed. V.C. Müller, Springer-Verlag, Berlin 2013, pp. 389–390.

[17] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, European Union 2019, pp. 12–13. This document is cited in later documents on AI, including the documents related to the Artificial Intelligence Act proposal that

was drawn from the EU Charter of Fundamental Rights and later expanded in the finalization of the Artificial Intelligence Act proposal. Ben Wagner explains that "EU fundamental rights are not understood as fundamental rights but rather as ethical imperatives to be complied with in a non-binding fashion."[18] In fact:

> In this sense these are "potential fundamental rights," developed under the shadow of hierarchy of the European Commission. They certainly cannot be claimed at present and if these potential fundamental rights are "violated" (whatever that means in the context of ethical commitments to uphold fundamental rights) they would be no legal recourse of any kind available. Indeed, it is in fact likely that these rights would actively need to be violated frequently and these violations would need to be made public widely, in order for the European Commission to be willing to do anything about their actual violation.[19]

However, to define "ethical AI", the EC created a single concept – *trustworthy AI* – composed of three distinct phenomena – law (the AI has to be law-abiding), ethics (the AI has to follow a set of action-guiding principles), and technics (the AI has to be robust). By doing so, the EC's proposal merged law and ethics into a single entity, even though it itself differentiates between law and ethics like Yampolskiy does, by strongly focusing on AI system *solution* via *value alignment* and in that way, at least on the surface, further attempted to prepare ground for subduing the environment which creates AI systems and the actor network that uses AI systems, doing so outwardly, that is, making the solution itself the starting point, thus going beyond "regulation by design."[20] For example, when the proposal states that an AI system should be "transparent," it means that all human and non-human elements in its entire life cycle have to align to the value of traceability and explainability[21] for it to successfully retain the accepted property, and so practices such

was formally adopted in June 2023. For a review, see L.A. DiMatteo, *Artificial Intelligence: The Promise of Disruption*, in: *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*, eds. L.A. DiMatteo, C. Poncibò, M. Cannarsa, Cambridge University Press, Cambridge 2022, pp. 12–14.

18 B. Wagner, *Ethics as an Escape from Regulation: From "Ethics Washing" to Ethics-Shopping?*, in: *Being Profiled: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen*, eds. İ.E. Bayamlıoğlu et al., Amsterdam University Press, Amsterdam 2018, p. 85.

19 Ibid.

20 L.A. DiMatteo, *Artificial Intelligence*, op. cit., p. 14.

21 European Parliament, P9_TA(2023)0236: Artificial Intelligence Act, Amendment 213, Article 4 a (p. 127).

as psychological targeting,[22] fake news, hate generation, preference recognition,[23] etc., should be considered for prohibition. If this conception was to be enforced with strict regulation and the appropriate administrative support, it might be representative of how potentially dangerous and exploitative new technologies could be used where they benefit humanity by aiming to create systems based on technical invention. It could be understood as a way of addressing the general problem of technical inventions taking control of social processes.

The idea of calling AI systems "ethical" further stems from the development of AI systems that exhibit autonomous behaviour, thus resembling a subject. This is highly debatable because most of what is considered "autonomous" in discussions on AI is most likely a more complex form of automation. In the simplest terms, a system that was automatized is a system that will once initiated continuously carry out the specified task by itself until completed without deviation –

[22] "Recent research in the field of computational social sciences […] suggests that people's psychological profiles can be accurately predicted from the digital footprints they leave with every step they take online. For example, people's personality profiles have been predicted from personal websites, blogs, Twitter messages, Facebook profiles, and Instagram pictures. This form of psychological assessment from digital footprints makes it paramount to establish the extent to which behaviours of large groups of people can be influenced through the application of psychological mass persuasion – both in their own interest (e.g., by persuading them to eat healthier) and against their best interest (e.g., by persuading them to gamble)" – S.C. Matz et al., *Psychological Targeting as an Effective Approach to Digital Mass Persuasion*, "PNAS" 2017, Vol. 114, No. 48, p. 12714, https://doi.org/10.1073/pnas.1710966114.

[23] An extreme case of preference recognition is AI's ability to detect sexual orientation solely by observing facial images and with much higher accuracy than human beings. These particular results were published by Yilun Wang and Michal Kosinski (Stanford University) in 2018, in a paper titled *Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images*: "Their decision to do the study at all, despite the evident risk to people living in countries where homosexuality is illegal, is justified by the authors in terms of the fact that if it is possible, then it represents a risk and should be public" – A. Campolo, K. Crawford, *Enchanted Determinism: Power without Responsibility in Artificial Intelligence*, "Engaging Science, Technology, and Society" 2020, Vol. 6, p. 12. Cases related to facial recognition are especially troublesome because they are notorious for the lack of certain interpretability of how and why the results are generated. See L.D. Introna, D. Wood, *Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems*, "Surveillance and Society" 2004, Vol. 2, Nos. 2–3, pp. 177–198, especially pp. 183–184. Developers are struggling to this day to reduce the black-box effect. For example, Wang and Kosinski also did not know exactly why their AI system is able to detect sexual orientation, and this is also becoming the problem in analysis or understanding "whether each action is performed in a responsible or ethical manner" – I. Gabriel, *Artificial Intelligence, Values, and Alignment*, "Minds and Machines" 2020, Vol. 30, No. 3, p. 412, https://doi.org/10.1007/s11023-020-09539-2.

"the machine is on; it runs its course."[24] To differentiate from simple automated systems,[25] it can be said that an *autonomous system* is "a system situated in an environment that senses the environment and acts on it in pursuit of its own agenda, in such a way that its actions can influence what it later senses."[26] Moreover, the capabilities of "learning," "adaption," and "choice-making" are added to such systems, with some authors emphasizing that it is about objects having "unsupervised activity."[27] But all these notions, which we would usually apply to living beings – perception, learning, adaption, having an agenda, choice-making, etc. – do not really transfer to machines. They are an *artificial* resemblance of organic capabilities because they are neither equivalent to capabilities found in living organisms nor the way in which they manifest can be found in living organisms. The unnecessary humanization of machines is maybe best seen in the use of the notion of "own agenda" instead of "specified task defined by the external user." So when encountered in the discourse, phrases signifying human behaviour and capabilities should be thought of as technical terms derived from the original notion applicable to living beings because of their orientational value in the knowledge landscape. Likewise, "autonomous" could be understood as higher-order automation because there is nothing in autonomous AI processes that differs from the fundamental trait of being a system that is continuously carrying out specified tasks by itself until completed without deviation. For this reason they can be only thought of as implicit subjects and their "morality" is only *functional* at their best, "where the machines themselves have the capacity for assessing and responding to moral challenges,"[28] but they retain moral inacces-

[24] H.M. Roff, *Artificial Intelligence: Power to the People*, "Ethics and International Affairs" 2019, Vol. 33, No. 2, p. 128, https://doi.org/10.1017/S0892679419000121.

[25] A class of auto-initialization lower than automation is automatization. "Automatic systems, such as a toaster in the civilian world or, to use a military example, an explosive triggered by a tripwire, respond mechanistically to environmental inputs. Automated systems, by contrast, operate based on multiple pre-programmed logic steps" – M.C. Horowitz, *Artificial Intelligence, International Competition, and the Balance of Power*, "Texas National Security Review" 2018, Vol. 1, No. 3, p. 40.

[26] S. Franklin, *History, Motivations, and Core Themes*, in: *The Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish, W.M. Ramsey, Cambridge University Press, Cambridge 2014, p. 27. Cf. H.M. Roff, *Artificial Intelligence*, op. cit., pp. 129–130.

[27] C. Allen, W. Wallach, *Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?*, in: *Robot Ethics: The Ethical and Social Implications of Robotics*, eds. P. Lin, K. Abney, G.A. Bekey, The MIT Press, Cambridge, MA, 2012, p. 55. "Unsupervised" to, still, "execute tasks on the designer's behalf" – E. Alonso, *Actions and Agents*, in: *The Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish, W.M. Ramsey, Cambridge University Press, Cambridge 2014, p. 235.

[28] Ibid., p. 57.

sibility – they cannot know that their operations are "moral," and what is or is not a "moral challenge" is recognized by human beings, not autonomous AI systems.

"Ethical AI" is altogether a clumsy expression because it subsumes the multitude of meanings hidden under the abbreviation "AI" and perpetuates the modern trend of the technical connectivity of moral subjectivity to non-living, non-self-conscious objects via norms.[29] "Ethically aligned AI", as proposed by IEEE Global Initiative,[30] is a better expression because it tells us that AI was aligned by something to mediate conduct towards itself and others without itself being a moral subject. The expression "ethical AI", although not to my scientific liking, is, however, pragmatic and applied widely. It should first be understood in the broadest sense as an AI system that, by its very existence, embodies preferred principles related to optimal moral behaviour in the human sense. Ethical AI is thus an AI system whose construction and performance is subject to predefined norms and values that are considered socially acceptable. However, what is "socially acceptable" in its universality is challenged by the realism of cultural relativism and personal preferences. "Ethical AI" as a term hides its structural complexity essentially related to the *ideological* component by which the social acceptability inherent to the term is limited.

In this paper, *ideology* is understood as "systematized ideas that, if followed in a prescribed manner, will lead to a preferred social outcome."[31] The preference of social outcome may aim at its possible universality, but it may not. In an armed conflict between two states, nations, ethnic groups, tribes, etc., social preferences are clashed despite some of them being possibly compatible. The existence of different cultural set-ups that generate different social preferences, for example, the acceptability of death penalty for apostasy, tells us that there are only two ways of developing and applying ethically aligned AI, either with the aim of supporting

[29] In philosophy this is usual for American and Dutch new waves of the philosophy of technology, and Luciano Floridi's circle of influence.

[30] IEEE Global Initiative, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 2019, URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

[31] N. Chitty, S. Dias, *Artificial Intelligence, Soft Power and Social Transformation*, "Journal of Content, Community and Communication" 2017, Vol. 6, No. 3, p. 1. Of course, there is a plethora of slightly different perspectives and uses of the concept of ideology, and the use of a more prominent approach, such as that of Karl Mannheim, Karl Marx, Marxists, Karl Jaspers, Herbert Marcuse, Jürgen Habermas, David Bloor or Michel Foucault, would certainly be useful for the analysis. However, Chitty and Dias's formulation is very effective for the discourse on ethical AI, especially since of those who refer to the concept of ideology at all in their work on AI, the majority of authors who have mentioned it use it without meaningful relevance to the general research on AI.

universally acceptable social preferences or otherwise. The latter is usually a sign of ideology building the foundation for a particular action. What this paper proceeds to show is that "ethically aligned AI," albeit discussed as if it is a matter of universal moral code, in practice embodies systematized ideas for a preferred non-universal outcome that is presented as an ethically aligned product. "Ideological limits" emanate from the core system of ideas embodied in the product or its application, in that any "ethical alignment" – either as engineered or applied – becomes a set of non-universal preferences that benefit some, but not all. In that sense, "ethical" becomes a simple descriptive term for *having a set of dispositional principles for expected conduct*, not a term referring to what is truly right or wrong, good or evil, morally permissible or impermissible, that we may further find to be universal or universalizable. The "ideological limit" thus denotes a boundary beyond which "ethical" is just a preference construct for a restricted gain.

The following section categorizes the dimensions of "ethical AI" and discusses the details of difference among them. The classification serves to show different ways of how the supposed ethical alignment can be carried out and how it relates to a difference between engineering practice, legal compliance, and social acceptability, for the purpose of showing how various instances of the problem of ideology come to the fore. These instances are then exemplified and discussed in the third section.

## 2. Classifying the Ethical in Artificial Intelligence

### 2.1. Basic Distinctions

To successfully tunnel through the ethical AI systems problem network, the simplest approach is to separate the presumed content into fundamental categories:

- ethical design of AI systems
- ethical development of AI systems
- ethical behaviour of AI systems
  - non-autonomous
  - autonomous
  - self-aware non-autonomous
  - self-aware autonomous
- ethical use of AI systems.

An AI system can be designed so that, for example, all information and actions related to the activity of the AI system are transparent and accessible/readable by anyone who has elementary information and digital literacy skills. This does not mean that the high-value data was properly tested when it was developed, and if it was, it does not mean that it was obtained in a fair way or without exploiting intellectual property loopholes. We can have an ethical AI compliant with current legal systems that appears socially acceptable, but was developed unethically. Even if the data was properly tested and obtained in a fair way, it still does not mean that the AI system was trained or managed ethically. One paradigmatic example is the functioning of OpenAI, a company that developed a sensible, amusing and reasonably useful application, ChatGPT, by using cheap labour,[32] switching from non-profit organization to profit company after achieving its developmental goal on the basis of donations,[33] and exploiting the uncontrolled data flow of the entire accessible Internet, including collective non-profit common-good efforts such as Wikipedia, to build its database for "training" AI for a service that then became privileged and now consumes 500 millilitres of water per 5 to 50 queries and spends an energy equivalent of up to 33,000 households per day.[34] The product may appear socially acceptable, it may offer clean and valuable data, but its developers may have exploited legal loopholes and weak links in the social environment for the product to become possible and feasible. The case is akin to enjoying an Apple smartphone that contains cobalt obtained through child labour in Congo mines.

---

[32] B. Perrigo, *OpenAI Used Kenyan Workers on Less Than $2 per Hour to Make ChatGPT Less Toxic*, "Time," 18.01.2023, URL: https://time.com/6247678/openai-chatgpt-kenya-workers/.

[33] C. Nduka, *How OpenAI Transitioned from a Nonprofit to a $29B For-Profit Company*, Hackernoon, 28.03.2023, URL: https://hackernoon.com/how-openai-transitioned-from-a-nonprofit-to-a-$29b-for-profit-company.

[34] Water consumption estimates were pre-reported in C. Novo, *The Water Cost of Artificial Intelligence Technology*, "SmartWaterMagazine," 12.09.2023, URL: https://smartwatermagazine.com/news/smart-water-magazine/water-cost-artificial-intelligence-technology. For a broader survey on AI's background water footprint, see the paper the report is based on: P. Li et al., *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*, arXiv:2304.03271 [cs.LG], https://doi.org/10.48550/arXiv.2304.03271. Energy estimates were a result of Sajjad Moazeni's research; basic information can be found in S. McQuate, *UW Researcher Discusses Just How Much Energy ChatGPT Uses*, University of Washington, 27.07.2023, URL: https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/. For an unrelated study on the growing energy footprint of AI, see A. de Vries, *The Growing Energy Footprint of Artificial Intelligence*, "Joule" 2023, Vol. 7, No. 10, pp. 2191–2194, https://doi.org/10.1016/j.joule.2023.09.004.

Similarly, an AI system may be both designed and developed in accordance with the expected conduct, that is, "ethically," but depending on what the AI system actually is, how well it is designed and developed, and how its use is regulated and limited, its ethical design and development may be denied in practice. A non-autonomous AI system, for instance, dialogic software such as ChatGPT, may provide dangerously inaccurate information about, for example, human conflict history or social status, regardless of the developer's best possible intentions, and could offer wrongful guidance in conduct to those who might ask for such a thing.[35] An autonomous AI system, such as the one implemented in an armoured combat vehicle, can be damaged, hacked or corrupted during war. The result can be an environmental miscalculation resulting in underage civilian casualties through action independent of human guidance. AI systems applied in predictive policing have already showed disastrous results because they are biased, racial, suggest oppressive monitoring practices and hamper elementary human rights.[36] Self-aware autonomous AI systems, which are currently only speculated about, have the same potential range of possible ethical misconduct as humans.

Ultimately, if an AI system were designed and developed in complete compliance with expected ethics and "behaved" accordingly, it could still be misused and exploited for unethical purposes. Unethical use must not be conflated with ethical AI, but the distinction still has to be made. For example, an AI system can be developed to simply track, record and analyze the movements of life systems. Such a system could be used to track animal populations in an ecosystem to help preserve biodiversity. But it can also be used to track undesirables, as in the two high-profile African cases where the Chinese company Huawei assisted the Ugandan and Zambian governments in tracking political opponents by sell-

[35] For example, in March 2023, the Belgian daily newspaper *La Libre* reported that a man had allegedly committed suicide after continuously exchanging information with an AI chatbot on an app called Chai. The man had previously been "increasingly pessimistic about the effects of global warming" and had isolated himself from family and friends in the pursuit of understanding the problem through the use of the dialogical AI system. See C. Xiang, *"He Would Still Be Here": Man Dies by Suicide after Talking with AI Chatbot, Widow Says*, Vice, 30.03.2023, URL: https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

[36] For an overview, see Fair Trials, *Automating Injustice: The Use of Artificial Intelligence and Automated Decision-Making Systems in Criminal Justice in Europe*, 9.09.2021, URL: https://www.fairtrials.org/articles/publications/automating-injustice/.

ing them AI-based equipment.[37] Here, too, is where ideological limits to ethical AI can be considered. From the perspective of EU citizens these practices may be considered socially unacceptable and legally questionable. Yet they did become a social reality for Uganda and Zambia, with whom we are connected at least through accepting Huawei products in our local stores and buying them for our business and amusement, and the legal system in Uganda and Zambia can support that kind of technological use. From the perspective of the upholder of the current state of affairs, neither the AI systems are unethical nor their use is unethical because fighting against the government is viewed as unethical. In addition, in the case of self-aware AI systems, even if everything to do with design, development, behaviour and use is ethically formidable, the instrumentalization of a self-aware entity is at least morally questionable, especially if such a system begins to pursue on its own an end that deviates from the intended means.

## 2.2. The Forness of Artificial Intelligence Systems

The stratification of ethical phenomena related to AI systems stems from the nature of AI systems as made things. Firstly, AI, narrowly understood as a study field of computer science and engineering,[38] broadly being a "wide range of technologies or an abstract large-scale phenomenon,"[39] is essentially an imitative solution[40] that becomes implemented into machine systems performing actions

[37] J. Parkinson, N. Bariyo, J. Chin, *Huawei Technicians Helped African Governments Spy on Political Opponents*, "Wall Street Journal," 15.08.2019, URL: https://www.wsj.com/articles/huawei-technicians-helped-african-governments-spy-on-political-opponents-11565793017#comments_sector. Uganda and Zambia belong to the top third of most corrupt countries in the world, as established by Corruption Perceptions Index.

[38] S. Franklin, *History, Motivations, and Core Themes*, op. cit., p. 15; S.M. Liao, *A Short Introduction to the Ethics of Artificial Intelligence*, in: *Ethics of Artificial Intelligence*, ed. S.M. Liao, Oxford University Press, Oxford 2020, p. 3. A variant definition to AI as discipline was given by Iason Gabriel as "the design of artificial agents that perceive their environment and make decisions to maximise the chances of achieving a goal" – I. Gabriel, *Artificial Intelligence, Values, and Alignment*, op. cit., p. 412.

[39] T. Hagendorff, *The Ethics of AI Ethics*, op. cit., p. 111.

[40] Authors discussing this variously refer to mimicry, imitation, and simulation. These are not entirely precise terms, but they are applicable to different contexts. Mimicry can certainly be applied to end products that have been biomorphized to be more accessible and user-friendly, or to actions that exhibit behaviour derived from the function of mimicry. AI systems certainly simulate in the broader sense of the word (originally, simulation referred to the creation of running models for the purpose of predicting its outcomes or representing it), but nevertheless they are based not only on trying to duplicate behaviour and outcomes, but also on duplicating

that use the distinct computation method resembling thought-processing, and appears as a material (physical/digital) entity performing tasks translated into understandable output through the computer interface and hardware shell.[41] It is a process that "exploits" a "realisation that nature, or human nature, works a certain way,"[42] constructed into material systems that combine multiple natural "effects" into a "chain of effects"[43] to our expected working advantage. Because AI was invented, designed, developed, and deployed by human beings, it should be understood as a *product* – a non-living produce of human beings. Being a software in a hardware shell, as products AI systems have all the characteristics of constructed artefacts – "object made by a human being that is not naturally present but occurs as a result of the preparative or investigative procedure by human beings."[44] For such an object to be, matter is "transformed such that the resulting physical construction has certain capacities or shows a particular kind of behaviour,"[45] attaining the status of objects that have a specific "practical 'for-ness,'"[46] which is generated by human activity. The element of *forness* explicates the fundamental aspect of artefacts as human-made *conveyants*. Being "for something" means that there is an interactor that will activate properties of conveyance in a specific artefact and cause an effect manifesting within the artefact and to its environment. For that reason, scholars in the second half of

certain internal processes or abilities of living beings, which would also make them emulative systems. However, all three concepts serve the purpose of imitation for a specific purpose. The aspect of imitation is important for understanding AI in relation to the general forms of machine learning today. Although most AI systems today use machine learning, imitation can also be achieved through other means, such as programmed execution rules masquerading as intelligent behaviour, as was the case with so-called expert systems in the 1980s. Machine learning by itself is "creation of software-based algorithms that build a mathematical model based on data, that can make decisions, predictions or perform tasks without being specifically programmed to do these tasks," usually attributed to AI (H. Seaton, *The Construction Technology Handbook*, John Wiley & Sons, Hoboken 2021, p. 102). This means that AI is an abstract idea, currently based only on machine learning, but it need not be so. It can be seen as a deployable capability to imitate for a specific purpose. The more complex the solution (expert system vs humanoid robot for elderly care), the clearer the attribute of imitation.

41 We should not rule out the possibility that AI systems in the future will not be computer based, which will certainly further blur the line between artificial and natural agency.

42 H. Seaton, *The Construction Technology Handbook*, op. cit., pp. 2–3.

43 Ibid., p. 4.

44 P.E. Ekmekci, B. Arda, *Artificial Intelligence and Bioethics*, Springer Nature Switzerland, Cham 2020, p. 17.

45 P. Kroes, *Technical Artefacts: Creations of Mind and Matter*, Springer, Dordrecht 2012, p. 3.

46 Ibid., p. 4.

the 20th century slowly began to conceptualize human products – technical artefacts foremostly – as mediators.

> When a technological artefact is used, it facilitates people's involvement with reality, and in doing so it coshapes how humans can be present in their world and their world for them. In this sense, things-in-use can be understood as mediators of human-world relationships. Technological artefacts are not neutral intermediaries but actively coshape people's being in the world: their perceptions and actions, experience and existence.[47]

As such, they may "ascribe new value to human beings, nonhuman things, and even to 'non-things' like future people and animals."[48] An AI system must be understood as a mediating technical product so that we can observe how its manifestation passes through phases of operational dimensions and comes into contact with the human lifeworld in which people articulate their preferred environment, for example, warfare and the promotion of political exceptionalism versus peace mediation and cosmopolitanism. This mediation of value grants them "normative power"; they are "examples of how code is law as well as how code creates law, or rather produces norms,"[49] which can be demonstrated by any number of applications, from the norm the AI imposes in filtering out discussions on social networks, through influence in medical or legal analysis and choice-making, to the selection of feasible workers, mortgages, or the creation of "new rules of interaction between economic agents" to "create a new form of

---

47 P.-P. Verbeek, *Moralizing Technology: Understanding and Designing the Morality of Things*, The University of Chicago Press, Chicago 2011, pp. 7–8.

48 L. Magnani, *Morality in a Technological World: Knowledge as Duty*, Cambridge University Press, Cambridge 2007, p. 13. Magnani gives an example: "Think for a moment of cities with extensive, technologically advanced library systems in which books are safely housed and carefully maintained. In these same cities, however, are thousands of homeless human beings with neither shelter nor basic health care. Thinking about how we value the contents of our libraries can help us to reexamine how we treat the inhabitants of our cities, and in this way, the simple book can serve as a moral mediator."

49 G. de Gregorio, *The Normative Power of Artificial Intelligence*, "Indiana Journal of Global Legal Studies" 2023, Vol. 55, p. 3, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4436287. Cf. L.A. DiMatteo, *Artificial Intelligence: The Promise*, op. cit., p. 11: "Lawrence Lessig has argued that coders and software programmers, by making a choice about the working and structure of IT networks and the applications that run on them, create the rules under which the systems are governed. The coders therefore act as quasi-legislators. In other words, 'code is law' is a form of private sector regulation whereby technology is used to enforce the governing rules."

social order."[50] The ability to detect patterns or specifics unavailable to human beings carries the capacity for formulating norms because the computational result widens the perspective on reality. When viewed in the light of the Artificial Intelligence Act, AI systems are basically used as enhancers to the preferred norms and generators of new incentives. This phenomenon will become even more evident when (if) there will be an artificial general intelligence, as the object will begin to constitute norms for itself. Moreover, as products in the commercial sense and as tools of commercialization, AI systems acquire an additional dimension of use and mediation that must be taken into account, especially since AI became a symbol of the ongoing Fourth Industrial Revolution as the first such revolution originating from the private sector.[51] The market is the only playing field of the private sector, and "the profit motive ultimately drives markets."[52]

For example, a non-autonomous, non-self-aware AI system can be designed, developed and used in medicine in a completely ethical way to reduce tremor in Parkinson's disease patients, but what is the cost of restoring the person's quality of life? Is such conditioned use of AI ethical? Furthermore, an AI system may be entirely ethically designed and developed and used to monitor geographic movements for positive purposes, but a company may decide to capitalize on its product by selling it to parties who use it with the intention of harming people, regardless of the terms of trade. Furthermore, consider the imbalance between government and the population: data-collection technologies and AI systems used in the public sector, especially in government institutions, increase knowledge about the population and the ability to exercise power over the population, while the population knows less and less about the government's activities and is not granted any privileges to make its plans and activities a matter exempt from the law and kept secret from the public (for example, military operations, international negotiations, surveillance of public space, etc.). Benjamin Baez, following Grigori Perelman, puts it aptly:

---

50 Å. Melkevik, *The Internal Morality of Markets and Artificial Intelligence*, "AI and Ethics" 2023, Vol. 3, No. 1, p. 115, https://doi.org/10.1007/s43681-022-00151-x.

51 To the point where China has decided to change its usual state politics, and support non-state, private companies in developing AI. See W.A. Carter, W.D. Crumpler, *Smart Money on Chinese Advances in AI: A Report of the CSIS Technology Policy Program*, Center for Strategic and International Studies, Washington 2019, p. 5.

52 N. de Marcellis-Warin et al., *Artificial Intelligence and Consumer Manipulations*, op. cit., p. 261.

> [A]long with nation-states, large corporations enjoy great control over information "resources" (which include actual workers in the information economy, such as systems analysts, academics, etc.), and combined with the fact that these large corporations own formerly public resources because of privatization, and that media is increasingly becoming concentrated in these corporations, we certainly can say without qualification that the increasing centralization and monopolization of information is not overstating matters. What this means, as Perelman points out, is that in addition to withholding information from the public, the owners can also manipulate and censor information, distorting the public's understanding of situations, and making it more difficult for people to challenge what is happening to them (Perelman, 1998, p. 78).[53]

From AI being viewed as a mediating technical product, we can derive its nature, on the one hand, as a *weapon*, and on the other, as an *artificial agent*.[54]

Being a mediating technical product, an AI system is always a product for something, a tool. When used for offence or defence, it must be considered a weapon, not so much because of the possibility of incorporating it into other weapons, but rather because the wide range of AI systems can be weaponized. Weaponization of AI systems should not be understood narrowly in the sense that an AI system designed as a mere tool is converted exclusively into a weapon. AI systems can serve as a weapon and be a tool at the same time. For example, police personnel using AI surveillance systems may target or monitor specific groups to gain personal benefit, although the use is certainly monitored and restricted to some degree. In the context in which a tool is used offensively against a living being, it behaves like a weapon, whether or not this was intended and whether or not there is a direct physical interaction typical of classical weapons, since the end goal is to endanger life.

The higher order of AI system utilization is the deployment of artificial agents because, in addition to computing advantages that imitate reasoning, an AI sys-

[53] B. Baez, *Technologies of Government: Politics and Power in the "Information Age"*, Information Age Publishing, Charlotte 2014.

[54] Authors would usually use words such as *actor*, *agent*, *subject*, *operator*, etc. All these words imply a natural, self-conscious action with the capacity to perform an intended action. Nothing of the kind can be attributed to artificial beings. However, the word *agent* can also be used for things, but given the possible misunderstanding that arises from calling an AI system an agent, it might make sense to call it an *artificial agent*.

tem imitates the capabilities of living beings that we can use only when we change modality from having an inherent end to *being-for-something*. An AI system does so by having a certain degree of unpredictable outputting by which it repositions itself and its actions in the framework in which it interacts with the lifeworld, as if it were a living entity. Slavery, cannibalism, animal service, and animal industry are some of the extreme variants of denying inherent ends to living beings, and this will always result in exploiting the organism's capacity for forness (the possibility to do this comes from the mutual ontological characteristic of living and non-living things that they exist as things). With technological solutions in the form of artificial agents (AI programmes, robots, unmanned vehicles, etc.), the specific capabilities are utilizable without ever risking to treat beings with inherent ends wrongly; however, AI systems as artificial agents instead of mere tools (weapons) expand their ethical relevance by having the particular "freedom" to affect the constitution of the lifeworld and because imitation alters how the life-like object affects human beings.[55] Michael C. Horowitz warns, and he is right to do so, that "AI seems much more akin to the internal combustion engine or electricity than a weapon. It is an enabler, a general-purpose technology with a multitude of applications,"[56] but it is precisely the level at which it can be utilized as a weapon with the capacity for imitating the agency of living beings that makes it suddenly important to constrain it. Why and how it is being constrained, however, is what defines the limits to the ethical and thus provokes a question concerning the ideological dimension of what has been declared "ethical," thus seemingly universal.

The following section finalizes and exemplifies the argument that there are realistic limits to having a truly ethical AI, and that these limits are fundamentally of ideological nature that may not be trumped by the current collective efforts.

---

[55] Humans can develop non-fictional emotions towards things, and our tendency to personify is heightened in encounters with artificial agents, especially given the tendency of developers to anthropomorphize or biomorphize their form or behaviour. See J. Blatter, E. Weber-Guskar, *Fictional Emotions and Emotional Reactions to Social Robots as Depictions of Social Agents*, "Behavioral and Brain Sciences" 2023, Vol. 46, e24, https://doi.org/10.1017/S0140525X22001716; M. Scheutz, *The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots*, in: *Robot Ethics: The Ethical and Social Implications of Robotics*, eds. P. Lin, K. Abney, G.A. Bekey, The MIT Press, Cambridge, MA, 2012, pp. 211–214.

[56] M.C. Horowitz, *Artificial Intelligence*, op. cit., p. 39.

## 3. Ideological Limits to Ethical Artificial Intelligence

Regardless of the agency level of an AI system, *forness* is the attribute through which the realization of ethical AI outlines its limits. For the phenomenon of forness to manifest, an intervention in existing matter by a creator or repurpuser is required to physically or symbolically construct the object and "attach" an intention to it. It is an act, and as an act it is historically contextual: it has a cause, a reason and a purpose associated with its action in an existing cultural environment, and thus creates a direction that defines what the subjected thing will be used for in the lifeworld it will affect. Like any other technological invention, "AI development does not take place in a vacuum. The development and adoption of technology is always highly social and cultural, embedded within a rich network of human and non-human actors,"[57] and so forness is what can be monitored to reveal cultural forces that push technological solutions into motion. It is impossible not to have these elements playing a role because technical solutions do not happen outside of cultural networks. Any such network is a structural coupling of communicational systems[58] that achieve the overarching identity. Its internal consistency reveals ideology, systematized ideas that, if followed in a prescribed manner, will lead to a preferred social outcome. Through its applied forness, a technological solution can, therefore, mediate the system's congruency of behaviour to reaffirm or advance the particular social habitus.

Here, a discursive difference has to be established between ideology as present in the ethical set-up of AI, the "ethical AI", and the *ideology of AI*. The ideology of AI presupposes a systematized idea that AI systems will make the world a better place, will solve all our problems, will correct all our mistakes, will make us work less, will fulfil all our desires even before we feel them, etc.[59] This advertisement strategy, endorsed by the leading national, supranational and corporate entities, essentially depicts an image of human beings as irreparably erroneous entities that cannot be trusted and should be supplemented or replaced wherever possible to increase work efficiency. Bruce J. Berman put it aptly already in the 1990s, at

[57] L. Munn, *The Uselessness of AI Ethics*, op. cit., p. 870.

[58] A.T. Polcumpally, *Artificial Intelligence and Global Power Structure: Understanding through Luhmann's Systems Theory*, "AI & Society" 2022, Vol. 37, No. 4, pp. 1492–1493, https://doi.org/10.1007/s00146-021-012198.

[59] For an overview of ideological narratives, see L. Sias, *The Ideology of AI*, "Philosophy Today" 2021, Vol. 63, No. 3, pp. 505–522, https://doi.org/10.5840/philtoday2021514405.

a time when AI systems were a matter of science-fiction stories to the majority of the human population:

> The tendency of the AI information processing model of mind to denigrate human intellectual abilities results in what Roszak terms a "technological idolatry" that reifies the computer metaphor, generating "a haunting sense of human inadequacy and existential failure" and propagating a deference to computers "which human beings have never assumed with respect to any other technology of the past" (Roszak, 1986: 44–45). This reveals the ideological importance of AI in both legitimating and restructuring of capitalist society and generating a technological imperative requiring the installation and subordination of human labour to "intelligent" computers.[60]

Berman cited a number of influential sources showing how the possible business advantages of AI systems are related to the capitalist worldview. His findings are in agreement with a recent examination by Mikko Vesa and Janne Tienari, who demonstrate AI's elementary appeal to the "elites":

> Imagine the promise of intelligent agent programs: they never miss a detail, they never forget, and they are constantly vigilant. Nor do they (supposedly) engage in petty games nor discriminate. They appear superior in their rationality and efficiency. They do not have "agency" in any classical sense and, as a consequence, no agent-principal problems. These programs do what they are told. Only they do so a bit better every time and they transcend human capabilities in processing information many times over. Promises of superior performance or competitive advantage derived from such technologies tend to be an easy sell for decision-makers. As such, intelligent agent programs and algorithms become objects of desire in complex ways for the power elite in society. The way AI delivers competitive advantages allows for a reconfiguration of power relations. Beneath it all lies the radical promise of organizing and organizations free of human concerns and shortcomings. In effect, this creates the premise to view intelligent agent programs as perfect rational agents. However, this is largely an experiential state associated with the mastery of such code by those who control them. This promise of rationality easily positions any critique as romantic, old-fashioned, and irrational.[61]

[60] B.J. Berman, *Artificial Intelligence and the Ideology of Capitalist Reconstruction*, "AI & Society" 1992, Vol. 6, p. 111, https://doi.org/10.1007/BF02472776.

[61] M. Vesa, J. Tienari, *Artificial Intelligence and Rationalized Unaccountability: Ideology of the Elites?*, "Organization" 2022, Vol. 29, No. 6, p. 1136, https://doi.org/10.1177/1350508420963872.

Vesa and Tientari propose that "artificial intelligence functions as an ideology as it manufactures normative idea(l)s of social reality and turns these into self-evident features of discourse (Fairclough, 1989) through which we are (not) able to make sense of the world,"[62] and they attempt to explain how the approach to AI contributes to the problem of proper accountability in contemporary technology-saturated global society. The process of pushing the global civilization into an "ideological state in which power and control are exerted algorithmically" can be understood as a natural continuity of 20th-century processes initiated and organized by then-growing technocrats.[63] To give an example that helps us see beyond the danger of falling into conspiracy theories, a charter written and published by OpenAI states the following:

> OpenAI's mission is to ensure that artificial general intelligence (AGI) – by which we mean highly autonomous systems that outperform humans at most economically valuable work – benefits all of humanity.[64]

The phrase "highly autonomous systems that outperform humans at most economically valuable work" inherently implies AI's purpose, which, in turn, suggests the systematic restructuring of civilization in the context of wealth distribution. Recall that Microsoft invested $13 billion in the project behind this statement. Given the current influence of OpenAI, their mission statement confirms the sense of ideology that has been growing since the inception of technocrats. However, given the dangers of a biased and superficial approach to any examination of the clash of classes, this requires a separate analysis, and thus the outlined grand narrative is not explored further in this research. The focus is on how ideology finds its way within the ethical set-up of AI systems.

That being said, the Artificial Intelligence Act can be seen as a paradigmatic example of the systematized proscription of ideas anchored into a single phenomenon around which the phenomenon itself wants to confirm its culture. The adopted text (amend. 15, p. 9) clearly states that "development and use of ethically embedded artificial intelligence" will have to "respect Union values and the Charter." It is here that the basic ideological limit begins to show contours, because what is "ethical" is equated with "Union values." This kind of formulation demonstrates the approach to ethics as being a *preferred* set of norms, altogether

---

62 Ibid., p. 1140.

63 These processes were explained well by Maurice Duverger in 1972. See M. Duverger, *Janus. Les deux faces de l'Occident*, Fayard, Paris 1972, esp. pp. 135–247.

64 OpenAI, *OpenAI Charter*, 9.04.2018, URL: https://openai.com/charter.

rendering the "ethical" arbitrary. It defeats the idea of a *universal ethos* practically – regardless of how much EU may claim that its values have a universal reach – and transforms the original concept of ethical as universal for every human being into a technical term. It also denies *value pluralism* as the foundation for a constructive integration of conflicting cultures, given that the same "ethical AI" in China, United States, Russia, India, Saudi Arabia, Alphabet Inc., Microsoft or the OECD will have different elements bound to the central concept. Without finding a way to overcome all sets of norms with a universal proposal, "ethical AI" may only be culturally interiorized and always completely prone to change, while the international scene of AI systems interaction will provoke cultural conflict and encourage ethics washing.[65]

Two sources can help us understand that it is not about ethical norms but about political and economic survival: national strategies and the EU social restructuring plan. Not a single national strategy of the relevant powers outside the EU, such as the United States and China, emphasizes anything other than benefits for their national gain, which from the perspective of ethics can certainly be understood as a form of ethical egoism, but in the end the harm to others is expected for the benefit of the self-oriented entity. The EU, on the other hand, is already perceived by both the EU and other political forces as an entity losing influence in the world and taking a beating in the Fourth Industrial Revolution, feeling threatened by China in particular.[66] In a special report of the Joint Research Centre on the "European perspective" on AI, it is emphasized that AI can "stimulate productivity and prosperity and lead to active work until a later age,"[67] that data is the "lifeline of Europe," and that "opening access to data and building interactions among participants is key to succeeding,"[68] presumably in the successful implementation of AI across the supranational entity for the stability of influence. In the light of this commentary, it is important to highlight an EU

65 Cf. I. Gabriel, *Artificial Intelligence*, op. cit., p. 426.

66 Cf. HAI, Artificial Intelligence Index Report 2023, Stanford University – Human-Centered Artificial Intelligence, Stanford 2023, URL: https://aiindex.stanford.edu/report/; B. Fricke Artificial Intelligence, 5G and the Future Balance of Power, "Konrad-Adenauer-Stiftung" 2020, No. 379, p. 6; A.T. Polcumpally, Artificial Intelligence, op. cit., p. 1498; Joint Research Centre, China: Challenges and Prospects from and Industrial and Innovation Powerhouse, Publications Office of the European Union, Luxemburg 2019, especially pp. 10–11, 20, 22, 31, 43–45.

67 Joint Research Centre, *Artificial Intelligence: A European Perspective*, Publications Office of the European Union, Luxemburg 2018, p. 56.

68 Ibid., p. 103.

report on the future of work, which states that "the acquisition of knowledge only through formal education will not be enough to thrive in the constantly changing world, which calls for the implementation of a lifelong-learning approach," requires "the constant re- and upskilling of workers,"[69] and a focus on "nurturing non-cognitive skills" because it "is becoming increasingly important for individuals' success in the labour market."[70]

The concept of "ethical AI" can mask the real normative for which the foundation is being developed. In the case of the Artificial Intelligence Act, the aim is gaining advantage on the global "playing field" but there is also need for risk-mitigation mechanisms for its population and reputation, and ways to overcome European national differences, as "the application of AI is often hampered by very restricted privacy laws, which make big data difficult to access."[71] Thus, it seems that the EU's behaviour confirms Hannah Arendt's claim that the social realm "is the form in which the fact of mutual dependence for the sake of life and nothing else assumes public significance."[72] For the EU's survival plan on AI to make sense, it needs to develop a fully accessible, free-flowing network of data collection equal to the networks of the United States, China, India, Russia, Japan, Australia, and other competing singularized entities, which entails not only heightened intrusion and exchange of population data but also control of the future production of data, as envisioned by the reports. The fundamental problem is that the recent progress of AI systems is due to data collected by "privacy-invasive social media applications, smartphone apps, as well as Internet of Things devices with its countless sensors."[73] Enforced regulations that supposedly regulate such data collection processes basically make no difference in practice, except to the creator of an ideological framework. These are the long-standing ethical problems of the post-privacy society, including the social and environmental costs of systematic reform, which are equally ignored and obscured by the concept of trustworthy AI.[74]

---

69 Joint Research Centre, *The Changing Nature of Work and Skills in the Digital Age*, Publications Office of the European Union, Luxemburg 2019, p. 28.

70 Ibid., p. 40.

71 B. Fricke, *Artificial Intelligence*, op. cit., p. 5.

72 H. Arendt, *The Human Condition*, The Chicago University Press, Chicago 1998, p. 46.

73 T. Hagendorff, *The Ethics of AI Ethics*, op. cit., p. 110.

74 Cf. ibid., pp. 105, 110.

In the service of the ideological system, any "ethical AI" is further diminished by the global military rivalry in which it is already assumed that "AI will give those who are well-prepared an upper hand" because "the data will enable one to 'know one's enemy as well as one knows oneself' and gain the competitive advantage."[75] The situation is so obvious that international relations and warfare experts openly discuss viable possibilities:

> Wealthy, advanced economies that have high levels of capital but also have light labor costs or small populations – middle powers such as Australia, Canada, and many European countries – often face challenges in military recruiting. For these countries, technologies that allow them to substitute capital for labor are highly attractive. […] countries can take advantage of the intersection of AI and robotics to overcome the problems caused by a small population.[76]

This creates another layer of invisible ethical problems piling up behind the idea of "ethical AI," in the sense that the broader framework of warfare remains ethically acceptable and only within this framework will questions about ethical behaviour arise. As Elke Schwarz observes, the "underlying question shifts from whether it is ethical to kill, to whether machines would do the killing better than humans. […] the ethical task at hand is to kill better and more humanely."[77] Data collection falls into the same category, as AI systems will be used to exploit the gathered information against the human source. In addition, Hagendorff emphasized that "one risk of this rhetoric is that 'impediments' in the form of ethical considerations will be eliminated completely from research, development and implementation. AI research is not framed as a cooperative global project [regardless of the emphasis in strategies on global cooperation], but as a fierce competition."[78]

Moreover, due to the limited impact of arbitrary ethical standards, those systems that insist on thorough and strict adherence to complex rules, such as the EU, may lose out in the race to win because of the rules they establish, raising the question of the moral defensibility of setting up AI systems in rigorously ethical ways. From the perspective of the population shaped by the Fourth Industrial Revolution, in the logic of racing the "ethical AI" might appear "unethical." This

[75] B. Fricke, *Artificial Intelligence*, op. cit., p. 5.
[76] M.C. Horowitz, *Artificial Intelligence*, op. cit., p. 46.
[77] E. Schwarz, *Death Machines: The Ethics of Violent Technologies*, Manchester University Press, Manchester 2019, p. 165.
[78] T. Hagendorff, *The Ethics of AI Ethics*, op. cit., p. 107.

question is underpinned by another concept with which political and corporate entities try to profit from the development of AI and limit its ethics: "hampering" of progress. For example:

> the right to explanation in the GDPR will come at cost in the efficiency or efficacy of the AI systems in question: optimisation and efficiency will be partially sacrificed for increases in transparency and accountability. While this is unproblematic in itself, as critics of regulation like to point out, such initiatives decrease the competitiveness of such systems on the global market, thus diminishing their likely overall representation and impact at the global level.[79]

This issue is linked to another problem on the level of the ethically aligned design of AI – the fact that the ethical properties which we would like AIs to have, such as transparency and explainability, may, on the one hand, prevent the development of highly efficient AI systems that find correlations "in data too huge for human to assess,"[80] and, on the other, lessen the possibility of non-human "intelligent" behaviour leading to new discoveries.

The general framework of ideological limits applies to all the listed categories of ethical AI, but already at the level of design and implementation experts are familiar with the so-called *inclusive design paradox*, where "positively improving a system to include as many values as possible might negatively influence the overall application,"[81] creating too many competing principles for the AI system to resolve it appropriately for everyone. Joris Krijger called this the effect of *inter-principle tension*, "the challenge of implementing multiple values and principles in one design," to which he added *intra-principle tension*, "the challenge of translating a single normative principle (in)to a specific technological design."[82] His division can be updated with the notion of *extra-principle tension*, which can be understood as the challenge of resolving competing norms between what is included in the AI system and what has been excluded. An AI system which by necessity has to adhere to particular values will enforce the subjugation to these values in every situation in which it finds itself. Little is known about what hap-

[79] H.-Y. Liu, *The Power Structure of Artificial Intelligence*, "Law, Innovation and Technology" 2018, Vol. 10, No. 2, p. 206, https://doi.org/10.1080/17579961.2018.1527480.

[80] H.M. Roff, *Artificial Intelligence*, op. cit., p. 137.

[81] J. Krijger, *Enter the Metrics: Critical Theory and Organizational Operationalization of AI Ethics*, "AI & Society" 2022, Vol. 37, No. 4, p. 1432, https://doi.org/10.1007/s00146-021-01256-3.

[82] Ibid.

pens when ethical AI encounters different types of norms, and yet no thorough research has been conducted by policy makers working on strategies and regulation, while scholars have only began to explore in more depth how communication between AI systems in a saturated environment should be processed.

Of the major problems related to the ideological elements of "ethical AI," last but not least is the demographic structure of those involved in the development and discussion on AI, dominated by the white male population with some common characteristics related to the underlying cultural and, possibly, biological traits. The tech-culture toxicity goes beyond science and business solutions,[83] as AI systems are heavily present in the video-gaming industry in which the past ten years were abundant with sexual, racial, and exploitation scandals, as well as labour abuse, usually in the working environment of leading giants such as Activision Blizzard, CD Project Red, and Electronic Arts. It is a "culture known for the hypermasculine coder or 'brogrammer,'" where "60% of women reported unwanted sexual advances."[84] Recently, a class action lawsuit that "has accused a widely celebrated tech company of fostering racist conditions for years, including daily subjection to racial slurs, being assigned menial jobs in a segregated area of the factory, and being passed over in promotions for management."[85] This is the same social circle that systematically ignores the application of ethical principles, as pointed out in section 1. Male-normative values are most evident in the domain of ethical design and ethical development of AI, particularly in the male approach to understanding AI and solving problems. Classical empirical studies show that "women do not, as men typically do, address moral problems primarily through a 'calculating,' 'rational,' 'logic-oriented' ethics of justice, but rather interpret them within a wider framework of an 'empathic', 'emotion-oriented' ethics of care."[86] However:

> In AI ethics, technical artefacts are primarily seen as isolated entities that can be optimised by experts so as to find technical solutions for technical problems. What is often lacking is a consideration of the wider context and the comprehensive relationship networks in which technical systems are embed-

[83] For an abundance of examples and the history of this approach, see S. Watcher-Boettcher, *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, W.W. Norton & Company, New York 2017.

[84] L. Munn, *The Uselessness of AI Ethics*, op. cit., p. 871.

[85] Ibid.

[86] T. Hagendorff, *The Ethics of AI Ethics*, op. cit., p. 103.

> ded. In accordance with that, it turns out that precisely the reports of AI Now (Crawford et al. 2016, 2019; Whittaker et al. 2018; Campolo et al. 2017), an organization primarily led by women, do not conceive AI applications in isolation, but within a larger network of social and ecological dependencies and relationships (Crawford and Joler 2018), corresponding most closely with the ideas and tenets of an ethics of care (Held 2013).[87]

In order to understand the extent to which the systems actually contribute to improving the quality of life of the population, we need to pay attention to who exactly is developing the framework for the use of AI, what characteristics and problems a particular user group has, what kind of approach they have to their discoveries or inventions, and whether they published anything that may clearly explain their motives. For example, in the context of social instability in the United States in relation to policing and minority populations, AI prediction or identity recognition systems simply cannot be deployed as isolated technical support akin to patrol vehicles, security cameras or emergency call networks, unless there is a specific agenda to embolden the ongoing stratification, because they are based on data reflecting past social practices riddled with racial behaviour and corruption.

If we look at AI systems as conveying technological nodes within a social system, then we can identify the kinds of cultural contexts that are attached to them and the proscriptions they mediate, that is, the system of ideas they indirectly represent or endorse. We then come to understand how they can behave as springboards for aims beyond their internal ethical set-up. Following the differentiation given at the beginning of sections 2 and 3, it can be concluded that ideological elements that affect ethical set-ups in AI systems and thus, by generating ideological bias in realistic deployment, limit what the "ethical" can achieve, manifest at three levels:

- social framework, endorsing AI systems development and application, which can be studied to identify the broadest forces and the most important actors within each social subnetwork endorsing AI systems to determine why they are being forced upon the citizens and what general claims and arguments for these claims are attached to the agenda;
- engineering framework, which can be studied to identify what cultural values were endorsed or implemented under the presented set of principles, as is the case with the EU, which wants AI systems to embody "Union values"

[87] Ibid., pp. 103–104.

specifically, or OpenAI, which exploited legal loopholes and economic instability;
- use of AI systems, which does not constrain ethical AI internally but externally, and can thus be studied in comparison to what the AI system was designed for, to grasp the exploitation, corruption or deviation of its ethical set-up, which is oblivious to social context, such as the deployment of AI systems in warfare, policing, and legal disputes.

Any formulation of ethical guidelines implies communication with ideological frameworks. However, many participants in the field of AI overlook the presence of ideological elements during the development, deployment and use of AI systems, and the fact that ethical AI is ethical only insofar as what stands for "ethical" is either universally acceptable or does not attempt to push an agenda. The situation with AI development is quite the opposite – it has become entangled with the ideological framework stemming from political and economic interests, and the practices surrounding the concept of ethical AI already show that the concept can serve as a tool of manipulation. The infusion of ideological elements into ethical regulations, which will eventually align with legal systems and gain social acceptance, needs to be examined in more depth.

## 4. Conclusion

There may be ways to perfect the design, development and behaviour of AI systems that support humanity in its evolution of the humane and resemble acceptable moral behaviour or an appropriate universal code of conduct, but the deployment of AI systems is the dimension where their utility is encumbered by the broader ideological framework that arises from the cultural conflicts and habits that have historically been in place. The exploration, discussion and development of "true" ethical AI is what would inevitably "hinder" the developmental progress of AI systems, as it would "impede" the particularist and exceptionalist political and economic agenda, which is certainly one of the reasons why the issue is systematically avoided or overlooked. But the problems I have highlighted and discussed in this paper, undoubtedly not the only ones plaguing AI, will persist alongside everything that will unfold with AI systems in the near future. In terms of how AI could genuinely contribute to humanity, "business and politics as usual" is the worst realization of its potential because it fails to address the

stalled course of civilizational development encumbered by conflicts, low quality of life, and resource depletion.

The main issue of justifying the fundamental idea that the use of structurally saturated AI systems is imperative for the "better future of humanity" is argumentatively buried by the positivist and pragmatic approach to AI, and that too is ideological in nature: there is a lacuna between using AI systems to increase the efficiency of personal endeavours or prevent harm, and building a global infrastructure to monitor the conversion of each of our atoms into fuel for the survival of the current framework. In order for us to understand what the "ethical" in ethical AI presupposed, it has to be deconstructed to its fundamental components. For example, the EU's Artificial Intelligence Act continuously emphasizes that AI systems have to have an ethical set-up and have to be regulated by law but these constraints should not hamper their development. This means that in any arbitrarily evaluated moral dilemma, the developmental breakthrough always prevails over the moral constraints. Thus, for example, if achieving the EU's goals in the AI race necessitates gathering extensive information about citizens and a systematic restructuring of their lives, we can expect that the EU will bypass inconvenient regulations, such as privacy laws, and trample over the unregulated. However, it will be doing so to perpetuate the existing political and economic system – a system that Europeans themselves shaped over the past 200 years – and not to change the course of European citizens' existence and foster a better approach to life. Because AI systems appeared in an epoch of major ideological conflicts, as technological inventions they must be interpreted as the possible mediators of ideological goals, meaning that the content of the ethical in "ethical AI" has its boundaries drawn by ideological elements defining the product.

The considerations presented in this paper were limited to drawing attention to the ways in which ideological elements can enter the ethical set-up, which, based on its name alone, is often misguidedly represented or thought of as universal. I aimed to show that even a quite transparent, straightforward approach to presenting ethical AI, such as that of the EU in the Artificial Intelligence Act, presupposes aims and limitations to its applicability that subdue the ethical principles selected to form the ethical set-up of AI. These aims and limitations belong to the broader ideological frameworks that become attached to AI systems. Further steps that can be taken to broaden the research are a closer inspection

of how ideological elements come into play at each designated level, followed by sequential case studies, and an attempt to develop and demonstrate a toolkit for identifying ideological bias.

## Bibliography

Allen C., Wallach W., *Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?*, in: *Robot Ethics: The Ethical and Social Implications of Robotics*, eds. P. Lin, K. Abney, G.A. Bekey, The MIT Press, Cambridge, MA, 2012, pp. 55–67.

Alonso E., *Actions and Agents*, in: *The Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish, W.M. Ramsey, Cambridge University Press, Cambridge 2014, pp. 232–246.

Arendt H., *The Human Condition*, The Chicago University Press, Chicago 1998.

Baez B., *Technologies of Government: Politics and Power in the "Information Age"*, Information Age Publishing, Charlotte 2014.

Beck U., *Gegengifte. Die organisierte Unverantwortlichkeit*, Suhrkamp Verlag, Frankfurt am Main 1988.

Berman B.J., *Artificial Intelligence and the Ideology of Capitalist Reconstruction*, "AI & Society" 1992, Vol. 6, pp. 103–114, https://doi.org/10.1007/BF02472776.

Blatter J., Weber-Guskar E., *Fictional Emotions and Emotional Reactions to Social Robots as Depictions of Social Agents*, "Behavioral and Brain Sciences" 2023, Vol. 46, e24, https://doi.org/10.1017/S0140525X22001716.

Boddington P., *AI Ethics: A Textbook*, Springer, Singapore 2023.

Campolo C., Crawford K., *Enchanted Determinism: Power without Responsibility in Artificial Intelligence*, "Engaging Science, Technology, and Society" 2020, Vol. 6, pp. 1–19.

Carter W.A., Crumpler W.D., *Smart Money on Chinese Advances in AI: A Report of the CSIS Technology Policy Program*, Center for Strategic and International Studies, Washington 2019.

Chitty N., Dias S., *Artificial Intelligence, Soft Power and Social Transformation*, "Journal of Content, Community and Communication" 2017, Vol. 6, No. 3, pp. 1–14.

Council of the European Union, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts: Analysis of the Final Compromise Text with a View to Agreement*, no. Cion doc. 8115/21, Brussels, 26 January 2024.

De Gregorio G., *The Normative Power of Artificial Intelligence*, "Indiana Journal of Global Legal Studies" 2023, Vol. 55, pp. 1–19, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4436287.

DiMatteo L.A., *Artificial Intelligence: The Promise of Disruption*, in: *The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics*, eds. L.A. DiMatteo, C. Poncibò, M. Cannarsa, Cambridge University Press, Cambridge 2022, pp. 3–17.

Duverger M., *Janus. Les deux faces de l'Occident*, Fayard, Paris 1972.

Ekmekci P.E., Arda B., *Artificial Intelligence and Bioethics*, Springer Nature Switzerland, Cham 2020.

European Parliament, P9_TA(2023)0236: Artificial Intelligence Act, June 2023.

Fair Trials, *Automating Injustice: The Use of Artificial Intelligence and Automated Decision-Making Systems in Criminal Justice in Europe*, 9.09.2021, URL: https://www.fairtrials.org/articles/publications/automating-injustice/.

Franklin S., *History, Motivations, and Core Themes*, in: *The Cambridge Handbook of Artificial Intelligence*, eds. K. Frankish, W.M. Ramsey, Cambridge University Press, Cambridge 2014, pp. 15–33.

Fricke B., *Artificial Intelligence, 5G and the Future Balance of Power*, "Konrad-Adenauer-Stiftung" 2020, Vol. 379, pp. 1–9.

Gabriel I., *Artificial Intelligence, Values, and Alignment*, "Minds and Machines" 2020, Vol. 30, No. 3, pp. 411–437, https://doi.org/10.1007/s11023-020-09539-2.

Hagendorff T., *The Ethics of AI Ethics: An Evaluation of Guidelines*, "Minds and Machines" 2020, Vol. 30, No. 1, pp. 99–120, https://doi.org/10.1007/s11023-020-09517-8.

HAI, *Artificial Intelligence Index Report 2023*, Stanford University – Human-Centered Artificial Intelligence, Stanford 2023, URL: https://aiindex.stanford.edu/report/.

High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, European Union 2019.

Horowitz M.C., *Artificial Intelligence, International Competition, and the Balance of Power*, "Texas National Security Review" 2018, Vol. 1, No. 3, pp. 36–57.

IEEE Global Initiative, *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*, 2019, URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.

Introna L.D., Wood D., *Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems*, "Surveillance and Society" 2004, Vol. 2, No. 2–3, pp. 177–198.

Joint Research Centre, *Artificial Intelligence: A European Perspective*, Publications Office of the European Union, Luxemburg 2018.

Joint Research Centre, *China: Challenges and Prospects from an Industrial and Innovation Powerhouse*, Publications Office of the European Union, Luxemburg 2019.

Joint Research Centre, *The Changing Nature of Work and Skills in the Digital Age*, Publications Office of the European Union, Luxemburg 2019.

Krijger J., *Enter the Metrics: Critical Theory and Organizational Operationalization of AI Ethics*, "AI & Society" 2022, Vol. 37, No. 4, pp. 1427–1437, https://doi.org/10.1007/s00146-021-01256-3.

Kroes P., *Technical Artefacts: Creations of Mind and Matter*, Springer, Dordrecht 2012.

Li P., Lang J., Islam M.A., Ren S., *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*, arXiv:2304.03271 [cs.LG], https://doi.org/10.48550/arXiv.2304.03271.

Liao S.M., *A Short Introduction to the Ethics of Artificial Intelligence*, in: *Ethics of Artificial Intelligence*, ed. S.M. Liao, Oxford University Press, Oxford 2020, pp. 1–42.

Liboreiro J., *European Regulators Put Microsoft's $13 Billion Bet on OpenAI under Closer Scrutiny*, EuroNews, 9.01.2024, URL: https://www.euronews.com/myeurope/2024/01/09/european-regulators-put-microsofts-13-billion-bet-on-openai-under-closer-scrutiny.

Liu H.-Y., *The Power Structure of Artificial Intelligence*, "Law, Innovation and Technology" 2018, Vol. 19, No. 2, pp. 197–229, https://doi.org/10.1080/17579961.2018.1527480.

Loke S.W., *Designed to Cooperate: A Kant-Inspired Ethic of Machine-to-Machine Cooperation*, "AI and Ethics" 2022, Vol. 3, No. 3, pp. 991–996, https://doi.org/10.1007/s43681-022-00238-5.

Magnani L., *Morality in a Technological World: Knowledge as Duty*, Cambridge University Press, Cambridge 2007.

Marcellis-Warin N. de, Marty F., Thelisson E., Warin T., *Artificial Intelligence and Consumer Manipulations: From Consumer's Counter Algorithms to Firm's Self-Regulation Tools*, "AI and Ethics" 2022, Vol. 2, No. 2, pp. 259–268, https://doi.org/10.1007/s43681-022-00149-5.

Matz S.C., Kosinski M., Nave G., Stillwell D.J, *Psychological Targeting as an Effective Approach to Digital Mass Persuasion*, "PNAS" 2017, Vol. 114, No. 48, pp. 12714–12719, https://doi.org/10.1073/pnas.1710966114.

McQuate S., *UW Researcher Discusses Just How Much Energy ChatGPT Uses*, University of Washington, 27.07.2023, URL: https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/.

Melkevik Å., *The Internal Morality of Markets and Artificial Intelligence*, "AI and Ethics" 2023, Vol. 3, No. 1, pp. 113–122, https://doi.org/10.1007/s43681-022-00151-x.

Munn L., *The Uselessness of AI Ethics*, "AI and Ethics" 2023, Vol. 3, No. 3, pp. 869–877, https://doi.org/10.1007/s43681-022-00209-w.

Nduka C., *How OpenAI Transitioned from a Nonprofit to a $29B For-Profit Company*, Hackernoon, 28.03.2023, URL: https://hackernoon.com/how-openai-transitioned-from-a-nonprofit-to-a-$29b-for-profit-company.

Novo C., *The Water Cost of Artificial Intelligence Technology*, "SmartWaterMagazine," 12.09.2023, URL: https://smartwatermagazine.com/news/smart-water-magazine/water-cost-artificial-intelligence-technology.

Nyholm S., *The Ethics of Crashes with Self-Driving Cars: A Roadmap, I*, "Philosophy Compass" 2018, Vol. 13, No. 7, e12507, pp. 1–10, https://doi.org/10.1111/phc3.12507.

OpenAI, *OpenAI Charter*, 9.04.2018, URL: https://openai.com/charter.

Parkinson J., Bariyo N., Chin J., *Huawei Technicians Helped African Governments Spy on Political Opponents*, "Wall Street Journal," 15.08.2019, URL: https://www.wsj.com/articles/huawei-technicians-helped-african-governments-spy-on-political-opponents-11565793017#comments_sector.

Perrigo B., *OpenAI Used Kenyan Workers on Less Than $2 per Hour to Make ChatGPT Less Toxic*, "Time," 18.01.2023, URL: https://time.com/6247678/openai-chatgpt-kenya-workers/.

Polcumpally A.T., *Artificial Intelligence and Global Power Structure: Understanding through Luhmann's Systems Theory*, "AI & Society" 2022, Vol. 37, No. 4, pp. 1487–1503, https://doi.org/10.1007/s00146-021-012198.

Roff H.M., *Artificial Intelligence: Power to the People*, "Ethics and International Affairs" 2019, Vol. 33, No. 2, pp. 127–140, https://doi.org/10.1017/S0892679419000121.

Scheutz M., *The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots*, in: *Robot Ethics: The Ethical and Social Implications of Robotics*, eds. P. Lin, K. Abney, G.A. Bekey, The MIT Press, Cambridge, MA, 2012, pp. 205–221.

Schwarz E., *Death Machines: The Ethics of Violent Technologies*, Manchester University Press, Manchester 2019.

Seaton H., *The Construction Technology Handbook*, John Wiley & Sons, Hoboken 2021.

Sias L., *The Ideology of AI*, "Philosophy Today" 2021, Vol. 63, No. 3, pp. 505–522, https://doi.org/10.5840/philtoday2021514405.

Simonite T., *What Really Happened When Google Ousted Timnit Gebru*, Wired, 8.06.2021, URL: https://www.wired.com/story/google-timnit-gebru-ai-what-really-happened/.

UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris 2022.

Verbeek P.-P., *Moralizing Technology: Understanding and Designing the Morality of Things*, The University of Chicago Press, Chicago 2011.

Vesa M., Tienari J., *Artificial Intelligence and Rationalized Unaccountability: Ideology of the Elites?*, "Organization" 2022, Vol. 29, No. 6, pp. 1133–1145, https://doi.org/10.1177/1350508420963872.

Vries A. de, *The Growing Energy Footprint of Artificial Intelligence*, "Joule" 2023, Vol. 7, No. 10, pp. 2191–2194, https://doi.org/10.1016/j.joule.2023.09.004.

Wagner B., *Ethics as an Escape from Regulation: From "Ethics Washing" to Ethics-Shopping?*, in: *Being Profiled: Cogitas Ergo Sum. 10 Years of Profiling the European Citizen*, eds. İ.E. Bayamlıoğlu, I. Baraliuc, L. Janssens, M. Hildebrandt, Amsterdam University Press, Amsterdam 2018, pp. 84–89.

Watcher-Boettcher S., *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, W.W. Norton & Company, New York 2017.

Xiang C., *"He Would Still Be Here": Man Dies by Suicide after Talking with AI Chatbot, Widow Says*, Vice, 30.03.2023, URL: https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says.

Yampolskiy R.V., *Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach*, in: *Philosophy and Theory of Artificial Intelligence*, ed. V.C. Müller, Springer-Verlag, Berlin 2013, pp. 389–396.

Edukacja Filozoficzna
ISSN 0860-3839, eISSN 2956-8269
DOI: 10.14394/edufil.2025.0003
ORCID: 0000-0002-7723-7175
ORCID: 0000-0003-2478-6151
ORCID: 0000-0002-2117-8145
ORCID: 0000-0003-0147-7700

# Computational Analysis for Philosophical Education: A Case Study in AI Ethics

Alex Cline
(Queen Mary University of London)

Brian Ball
(Northeastern University London)

David Peter Wallis Freeborn
(Northeastern University London)

Alice C. Helliwell
(Northeastern University London)

Kevin Loi-Heng
(Northeastern University London)

**Abstract:** This paper explores what computational methodologies can tell us about philosophical education, particularly in the context of artificial intelligence (AI) ethics. Taking the readings on our AI ethics and responsible AI syllabi as a corpus of AI ethics literature, we conduct an analysis of the content of these courses through a variety of methods: word frequency analysis, term frequency–inverse document frequency (TF–IDF) scoring, document vectorization via SciBERT, clustering via *k*-means, and topic modelling using latent Dirichlet allocation (LDA). We reflect on the findings of these analyses, and more broadly on what computational approaches can offer to the practice of philosophical education. Finally, we compare our approach to previous computational approaches in philosophy, and more broadly in the digital humanities. This project offers a proof of concept for how contemporary natural-language processing techniques can be used to support philosophical pedagogy: not only to reflect critically on what we teach, but to discover new materials, explore conceptual gaps, and make our courses more accessible to students from a range of disciplinary backgrounds.

**Key words:** AI ethics, philosophy education, computational philosophy

## 1. Introduction

What can computational methods – particularly artificial intelligence (AI) – tell us about AI ethics education? In this paper, we apply computational approaches to interrogate our AI ethics courses. As philosophers working in the philosophy of AI, we are interested in what computational methods can add to philosophical studies, and vice versa.

In our philosophy and computer science programmes (MA Philosophy & Artificial Intelligence, MSc Artificial Intelligence & Ethics, BSc Philosophy and Computer Science), AI ethics education forms an integral part of our teaching. Our students have a wide range of backgrounds, though of course many have been trained in either philosophy or computer/data science. Our aim is to provide philosophical and computational education simultaneously, to equip students with the skills they need to responsibly engage with AI technology. Given this ethos, we have decided to apply computational methodologies to our own practice, by investigating some of the philosophy courses on these programmes. Our aim is to gain insight into our pedagogical approach and to develop a project which we can (hopefully) share with our students. In order to test our thought that computational tools can be useful for pedagogical and philosophical goals, we have conducted a computational analysis of the texts we set for students across two courses in AI ethics ("AI and Data Ethics" and "Advanced Topics in Responsible AI"). We have curated these papers over several years, and after completing both courses, we want our students to have covered a variety of classic and current topics in AI ethics and responsible AI. Having gathered the recommended readings for these courses, we utilized some standard Python-based natural language processing (NLP) techniques to analyse our corpus of texts.

In this paper, we explain our methodology and discuss the results of our analysis. We begin (section 2) with a description of the dataset, consisting of the reading materials assigned in two of our advanced (advanced undergraduate, MA and MSc level) philosophy courses on AI and data ethics, and explain how we prepared the texts for computational analysis. In section 3, we discuss the ethical considerations for this project. In section 4, we describe the NLP techniques we used to explore this corpus: from relatively simple tools, such as word frequency analysis and term frequency–inverse document frequency (TF–IDF) scoring, to more complex machine learning approaches, including document vectorization

via SciBERT, clustering via *k*-means, and topic modelling using latent Dirichlet allocation (LDA). Each of these methods offers a different lens through which to understand the themes of our syllabi. Word frequency and TF–IDF give us a surface-level, yet still informative, comparative view. SciBERT vectorization and clustering allow us to explore semantic relationships within the corpus. Topic modelling, finally, enables us to identify and interpret latent themes running throughout this body of literature.

Finally, in section 5, we discuss the broader implications of our approach, both for AI ethics education and for philosophy more generally. As philosophers working on AI, we see this project as a two-way exchange: using computational tools to enhance philosophy teaching and using philosophy to reflect critically on the use of such tools. We situate our work in the context of digital humanities, noting that while computational methods have been widely used in literature, history, and linguistics, they remain relatively underexplored in philosophy. This project offers a proof of concept for how contemporary NLP techniques can be used to support philosophical pedagogy: not only to reflect critically on what we teach, but to discover new materials, explore conceptual gaps, and make our courses more accessible to students from a range of disciplinary backgrounds. We conclude (section 6) with a call for further work in computational philosophy and philosophical pedagogy – and outline our plans for future analysis and engagement with students as collaborators in this ongoing exploration.

## 2. Dataset

The dataset utilized in this research consists of the required and supplementary readings assigned in two upper-level philosophy courses we teach: "AI and Data Ethics" and "Advanced Topics in Responsible AI." These are graduate-level or advanced undergraduate courses aimed at students from diverse disciplinary backgrounds, including philosophy, computer science, and the social sciences, as well as many graduate students with experience in industry. As instructors, we have curated the readings to offer both foundational and contemporary perspectives in the broad field of AI and data ethics. The goal is to introduce students to a wide range of normative concerns and philosophical methods, while also equipping them with the analytical tools to evaluate real-world technologies, applications

and policies. After completing the two courses, students should have established knowledge of essential topics in responsible AI and AI ethics, as well as the necessary skills to engage in normative discussions on emerging advances in AI.

The selected readings include a mix of philosophy papers, technical and policy-oriented research, and interdisciplinary contributions from fields such as computer science, law, economics and education. Authors in the corpus range from prominent philosophers to computer scientists discussing algorithmic bias, as well as economists, legal scholars writing on data privacy and AI regulation, and some technical practitioners of AI.

Together, the two courses span 22 weeks of teaching and 17 distinct thematic topics. Topics in course 1, "AI and Data Ethics," include:

- "What Is AI and Data Ethics?,"
- "Autonomous AI and Responsibility,"
- "Artificial Moral Agency,"
- "Personhood and Robot Rights,"
- "Algorithmic Bias and Fairness,"
- "Safe AI (Including Black Boxes, Transparency, and Explainability),"
- "Data, Democracy, and Misinformation,"
- "Privacy and GDPR,"
- "Superintelligence and the Control Problem,"
- "Regulation,"
- "Value Sensitive Design."

Topics in course 2, "Advanced Topics in Responsible AI", include:

- "What Is Responsible AI?,"
- "AI and Work,"
- "AI and the Creative Industries,"
- "AI and Education,"
- "AI and Human Interaction,"
- "AI and Sustainability."

From these topics, we collected the full set of assigned readings, resulting in a corpus of 184 distinct texts. These included journal articles, book chapters, and reports. From a technical perspective, we treated each reading as a single document in our corpus. The documents were compiled in plain text format.

Before we could move into computational analysis, we focused on preparing the textual data. Clean and standardized text is essential to ensure that any patterns we uncovered would be meaningful and as free from noise as possible. This

step sets the foundation for later stages of the project, including vectorization and clustering. Without a careful cleaning and normalization process, later stages, like similarity measurement or topic modelling, become vulnerable to distortion by irrelevant or redundant information. The SpaCy model was a useful, lightweight tool for our NLP tasks. The tool allowed us to tokenize the text into words and sentences, lemmatize words to their base forms, and remove punctuation and irrelevant characters. The aim here was to ensure that related terms – such as "machines" and "machine" – would be treated consistently.

## 3. Ethics

Several ethical issues were considered when conducting this analysis. We did not use any human subjects, and also did not utilize any personal information in our analysis, so human subject considerations were not applicable. The authors of the courses under analysis are all part of the project team and granted permission for their syllabi to be used for this analysis.

As we are interested explicitly in AI ethics in this paper, we also considered the ethics of the use of texts for analysis by AI. Whilst the texts in our corpus were all available online, and particularly for educational purposes, we have not made this corpus openly accessible in order to ensure we do not breach copyright protections. As we are utilizing AI to analyse our corpus of texts, we were also particularly aware of current debates in intellectual property and AI.[1] There is a growing debate around training data, reproduction, and attribution in the context of generative AI. However, the tools used in this project, including word frequency counters, TF–IDF models, SciBERT embeddings, and topic modelling, are all predictive rather than generative. As such, none of these methods produced new textual output derived from the source material; rather, we deployed these tools to extract patterns and representations from the existing dataset, in ways that are standard in computational linguistics and the digital humanities.

Finally, while student engagement is an important motivation for this work, no student data was utilized for this research.

[1] Our use of these texts falls under the scope of academic research and teaching, and we believe it is justified under principles of fair dealing, particularly given the non-commercial and scholarly nature of the work, in compliance with Section 29 of the UK Copyright, Designs and Patents Act 1988 (UK CDPA 1988, Section 29A – "Copies for text and data analysis for non-commercial research").

## 4. Analysis

### 4.1. Word Frequency Analysis

The first analytic tool we turned on our corpus of texts was a word frequency counter. This simple computational technique counts the number of times a word appears in a document, or collection of documents. This allowed us to identify the words that appear most frequently in our collection of papers, and produce the word cloud, where the most frequently used words appear largest in size, shown in Figure 1.

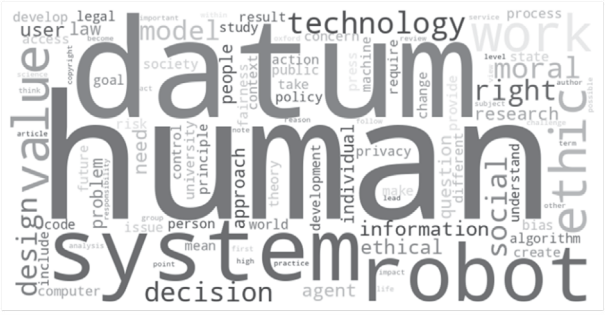



Figure 1. Word cloud of frequently appearing words in the corpus

We found that across our two courses, the highest frequency unique word used was "human." As our courses are primarily focussed on technology and ethics, this was perhaps somewhat surprising. However, it is likely that authors in our corpus are discussing humans in contrast with data (second most common) and machines (eighth most common), which are their explicit focuses. Words such as "work," "right," "value," "bias," and "ethic" are to be expected, given the topics in our courses. However, words such as "press," "social," "individual," "public," and "state" do not obviously correspond to particular topics and seem to highlight the social and societal focus of the courses.

Term frequency itself has limited utility for telling us about unique features of a corpus of texts. It could be, for example, that (contrary to the conjecture above) "human" is something that comes up in philosophical works in general. To find out more about the unique features of this body of texts, we conducted another analysis.

## 4.2. TF–IDF

To further examine whether our conjecture regarding word frequency was plausible, we decided to analyse word frequency further. We ran another measure on the corpus: a TF–IDF.[2] This NLP technique is typically used to evaluate the relative importance of a word in a document compared to its importance in the corpus as a whole. Rather than simply counting the frequency of use for each word, a TF–IDF can show which words are more common in our AI ethics corpus compared to a larger, or alternative, corpus of texts.[3]

Table 1. Top 10 words in each of the datasets

| AI ethics canon | Wittgenstein corpus |
|---|---|
| human | philosophy |
| ethic | Wittgenstein |
| moral | philosophical |
| robot | language |
| data | theory |
| system | political |
| technology | social |
| design | review |
| agent | science |
| develop | knowledge |

[2] K. Spärck Jones, *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, "Journal of Documentation" 1972, Vol. 28, No. 1, pp. 11–21, https://doi.org/10.1108/eb026526.

[3] Term frequency (TF) measures how often a term appears in a document relative to the total number of terms in that document. For a given term t_*j*, the term frequency is defined as TF_*j* = t_*j* / Σ t_*i*, where t_*j* is the number of times term *j* appears in the document, and Σ t_*i* is the total number of terms in the document. However, words that appear frequently across the entire corpus may be less informative. To account for this, inverse term frequency (IDF) is defined by, IDF_*j* = log(*N* / (1 + *n_j*)), where *N* is the total number of documents in the corpus, and *n_j* is the number of documents in which term t_*j* appears. The "+ 1" in the denominator avoids division by zero. We then define the TF–IDF score for term t_*j* as the product, TF–IDF_*j* = TF_*j* × IDF_*j*.

Of course, in this case we were not just interested here in individual papers, but the body of works as a whole. In order to complete a TF–IDF measure then, we required a contrasting corpus of texts. Following work from some members of our team on Wittgenstein and AI, we had a Wittgenstein corpus available;[4] a body of papers (accessed through JSTOR) discussing the work of Wittgenstein. This corpus, comprised of 64,000 total documents, was made on Constellate (from Ithaka), with their dataset builder from papers on JSTOR.[5]

When we compare these two analyses, we start to see the relative importance of these terms in the text. "Human," for example, is not just the most frequent unique word, but it is particularly important in the AI ethics papers compared to works discussing Wittgenstein. "Wittgenstein" is the second most important word in the Wittgenstein papers (a comforting sign that our analysis was working). Furthermore, in the Wittgenstein corpus, "philosophy" and "philosophical" are particularly prevalent. This may reflect the metaphilosophical nature of Wittgenstein's work (and thus discussions of his work) but may also reflect the relative lack of importance of "philosophy" in the AI ethics corpus, which spans more disciplines (such as law, computer science, and engineering).

### 4.3. Using AI: Vector Representations and Cosine Similarity

Few nowadays would consider the NLP techniques we have discussed so far to involve AI: in particular, the computational methods employed operate directly on textual data, here the full papers from our two course reading lists. Since research papers are written in natural language, they need to be converted into a numerical format that a computer can read and interpret if contemporary AI techniques are to be deployed on them. We did this using SciBERT,[6] a transformer model pre-trained on scientific texts based on the BERT model.[7] SciBERT converts each

[4] B. Ball, A.C. Helliwell, A. Rossi, *Wittgenstein and Artificial Intelligence: Mind and Language*, Anthem Press, London 2024; B. Ball, A.C. Helliwell, A. Rossi, *Wittgenstein and Artificial Intelligence: Values and Governance*, Anthem Press, London 2024.

[5] JSTOR Dataset ID: 77934734-096e-6982-c1de-af09599cd73e. Wittgenstein about philosophy – Applied philosophy, Philosophy – Axiology, Philosophy – Epistemology, Philosophy – Logic, Philosophy – Metaphilosophy, Philosophy – Metaphysics limited to document type(s) book, article from 1900–2023.

[6] I. Beltagy, K. Lo, A. Cohan, *SciBERT: A Pretrained Language Model for Scientific Text*, arXiv:1903.10676, https://doi.org/10.48550/arXiv.1903.10676.

[7] J. Devlin et al., *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805, https://doi.org/10.48550/arXiv.1810.04805. SciBERT is a pre-trained

document into a high-dimensional vector – essentially a mathematical "fingerprint" that captures the semantic content of the text. This allowed us to compare texts not by the words they contain directly, but by their learned representations: encodings that capture patterns of semantic meaning based on usage and context across the corpus.[8]

It is helpful to contrast this with another common technique in the digital humanities, which is to make use of Word2Vec.[9] Unlike SciBERT, which creates a single vector for an entire document, Word2Vec assigns vectors to individual words. A model is trained (actually a number of them) on a corpus, and this associates a vector – not with each document, as in our approach, but – with each word. The vector in question is used for next-word prediction: that is, the algorithm aims to associate a vector with each word that determines probabilities for the other words in the vocabulary that they occur next (in the corpus). Accordingly, each word's location in the vector space represents its usage (or distribution) within the corpus (that is, its associations with other words). This vindicates J.R. Frith's dictum, "You shall know a word by the company it keeps"[10] – and it allows us to compare, for example, the conceptualizations of words across corpuses.

After generating a vector for each document in our corpus using SciBERT, we computed pairwise cosine similarity scores between them.[11] A similarity score of

language model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically trained on scientific texts. In essence, this transforms raw text into numerical representations through a process known as contextual embedding, i.e., generating a vector for each token, based not just on the word itself, but on the surrounding words in both directions. Through sufficient training on a large sample, the model learns which words are most relevant to each other in context, even when those relationships are fairly weak, or the words are separated by long spans of text. For our analysis, we used the pooled output from SciBERT to produce a single-vector representation for each document. This vector can be understood as a dense, high-dimensional summary of the document's semantic context.

8 Y. Bengio et al., *A Neural Probabilistic Language Model*, "Journal of Machine Learning Research" 2003, Vol. 3, pp. 1137–1155.

9 T. Mikolov et al., *Efficient Estimation of Word Representations in Vector Space*, arXiv:1301.3781, https://doi.org/10.48550/arXiv.1301.3781; T. Mikolov et al., *Distributed Representations of Words and Phrases and Their Compositionality*, arXiv:1310.4546, https://doi.org/10.48550/arXiv.1310.4546.

10 J.R. Firth, *A Synopsis of Linguistic Theory 1930–1955*, in: *Studies in Linguistic Analysis*, Blackwell, Oxford 1957, pp. 1–32.

11 Cosine similarity measures the angle between two vectors in a high-dimensional space, given by their normalized dot product. The idea is simple: if two documents are represented by vectors

**1** indicates highly similar documents (identical in vector space), while a score near **0** indicates very different content. This allowed us to measure semantic similarity between papers, providing a foundation for a clustering analysis (see below).

### 4.4. Using AI: Clustering Papers into Meaningful Groups

We were also interested in drawing out where papers in our canon were grouped together around different subjects and themes. To examine this, we utilized a couple of methods. First, we applied *k*-means clustering, an unsupervised machine learning technique that groups papers into clusters based on their similarity.[12] It works on unlabelled data (that is, data without defined categories or groups). The algorithm first randomly selects central points, called centroids, then uses algorithms to automatically find common themes and structures in the data. We repeated the clustering with different *k* values to find different groupings. By experimenting with different *k* values we determined the best number of clusters. For this we used techniques like the elbow method and silhouette score to find a suitable number given the trade-off between better representing the data and using more clusters. We picked six clusters to move forwards.

We tested a range of values for *k*, the number of clusters, varying the number of clusters from 1 to 32, specifically testing *k* in [1, 2, 3, 4, 5, 6, 8, 12, 16, 20, 24, 28, 32]. For each clustering solution, we evaluated the results using the silhouette score (see Fig. 2),[13] a standard metric for assessing the quality of clustering.[14] The silhouette score captures both cohesion (how close each document is to the other documents in its cluster) and separation (how far it is from documents in other

that point in the same direction, they are semantically similar; if the vectors are orthogonal, then they are unrelated. Unlike the Euclidean distance, which measures how far apart two points are, cosine similarity focuses on the orientation of the vectors rather than their magnitude.

12 H. Steinhaus, *Sur la division des corps matériels en parties*, "Bulletin de l'Académie Polonaise des Sciences, Classe III" 1956, Vol. 4(12), pp. 801–804; J.B. MacQueen, *Some Methods for Classification and Analysis of Multivariate Observations*, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: *Statistics*, University of California Press, Berkeley–Los Angeles 1967, pp. 281–297; F. Pedregosa et al., *Scikit-Learn: Machine Learning in Python*, "Journal of Machine Learning Research" 2011, Vol. 12, pp. 2825–2830.

13 The silhouette score for a given document is calculated as $(b - a) / \max(a, b)$, where $a$ is the average distance to other points in the same cluster (i.e., intra-cluster cohesion), and $b$ is the average distance to points in the nearest neighbouring cluster (i.e., inter-cluster separation).

14 P.J. Rousseeuw, *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*, "Computational and Applied Mathematics" 1987, Vol. 20, pp. 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

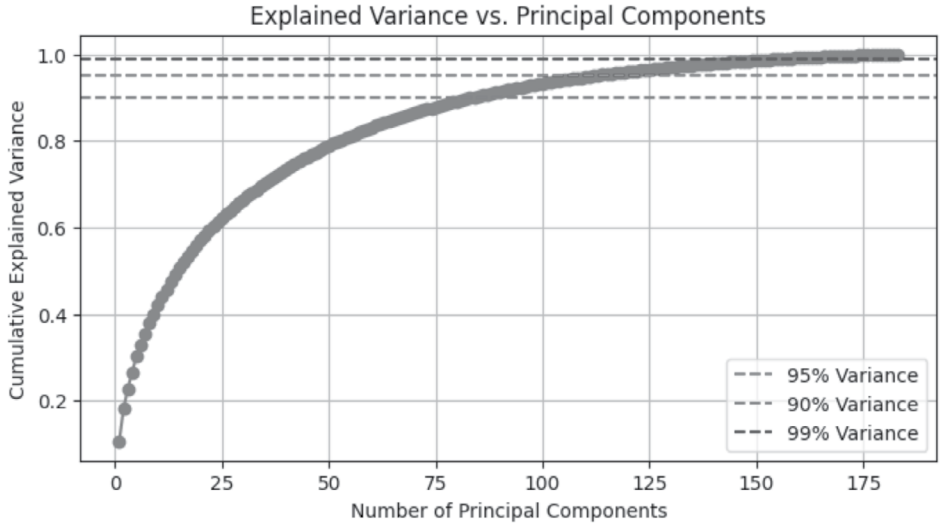

Figure 2. Silhouette score versus number of clusters

clusters). Scores range from –1 to 1, with higher values indicating more well-defined and internally coherent clusters.

After identifying candidate values of *k* that produced relatively high silhouette scores, we further examined the resulting clusters to evaluate their interpretability. This involved identifying central documents – those that were closest to the centroid of their cluster – as well as outlier documents that were located on the periphery of a cluster or between two clusters.

The *k*-means clustering algorithm is known to struggle with very high-dimensional data.[15] Since the SciBERT embeddings we used to represent each document exist in a fairly high, 768-dimensional space, we applied a dimensionality reduction technique to make the data more tractable for clustering. To do this, we used principal component analysis (PCA), a linear algebra-based method that transforms the original high-dimensional data into a lower-dimensional space

[15] This is an example of the so-called "curse of dimensionality." Distance metrics, as used for *k*-means clustering, become less informative as the number of dimensions increases. In such spaces, all points tend to become approximately equidistant from one another, making it difficult for the algorithm to identify meaningful groupings. Additionally, high-dimensional data tends to be sparse, which further reduces the effectiveness of clustering algorithms that assume dense, well-separated regions.




```
Number of components to retain 90% of variance: 84
Number of components to retain 95% of variance: 112
Number of components to retain 99% of variance: 157
```

Figure 3. Explained variance versus principal components

while preserving as much of the data's variance as possible. However, there is necessarily a trade-off between compressing the data and preserving the salient structural features. PCA works by identifying the orthogonal directions (called "principal components") along which the data varies the most and projecting the data onto a subset of those directions.[16] In our analysis, we chose to select the number of components such that 95% of the total variance in the original data was preserved (see Fig. 3). This corresponded to 112 principal components, which we used as the input space for the *k*-means clustering. The data is shown in Figure 4, classified into different numbers of clusters and then projected onto just two dimensions for visibility.

[16] More formally, PCA finds a new set of orthogonal axes – linear combinations of the original dimensions – ordered by the amount of variance in the data they explain. The first principal component captures the largest possible variance, the second captures the largest variance orthogonal to the first, and so on. By retaining only the top $N$ components, we reduce the dimensionality of the data while maintaining the majority of its informational structure; see I.T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, New York 2002.

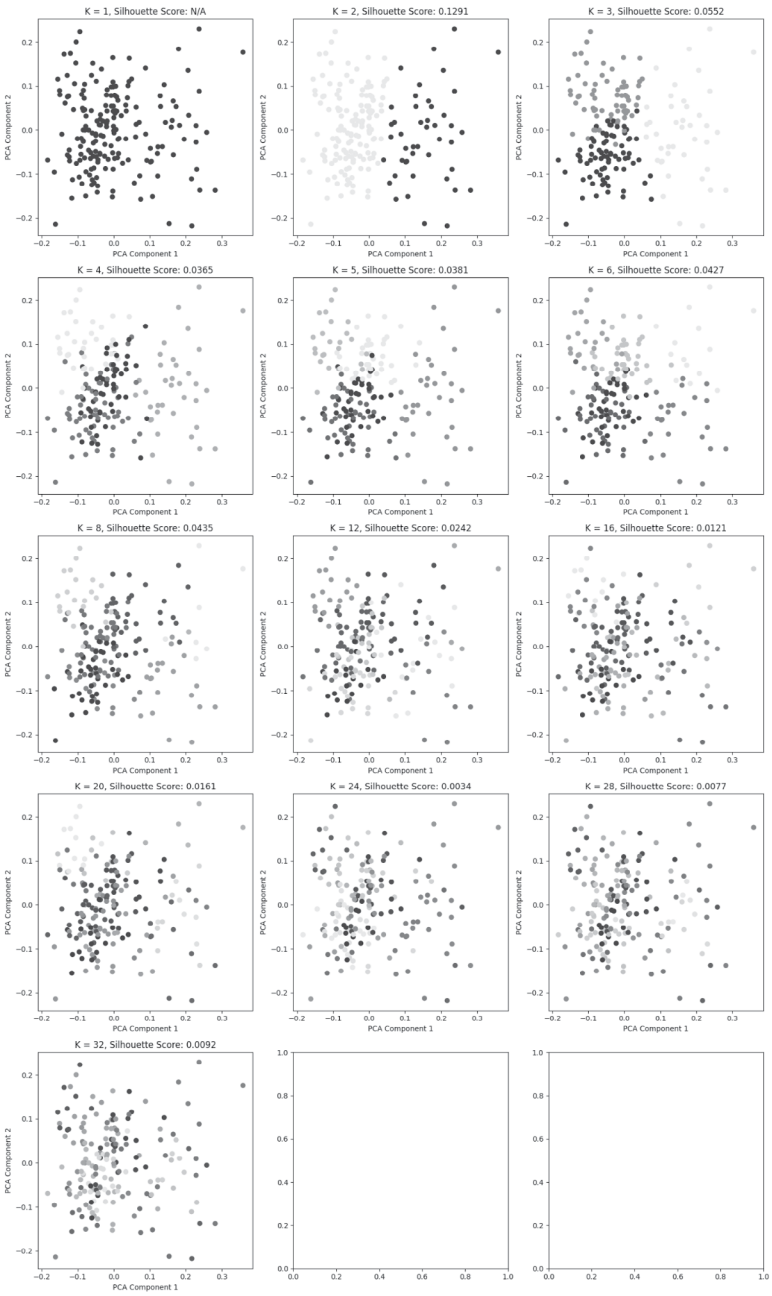



Figure 4. *K*-means clustering with principal component analysis showing the division of papers depending on different numbers of clusters. The dots represent the papers in the canon, with colours representing the clusters to which they belong

To ensure that the clustering results were meaningful, we checked whether each paper had the highest similarity to the average of its assigned cluster. The fact that 100% of papers were most similar to their own cluster's average reassured us that the model was making reasonable groupings.[17] In order to visualize these clusters, we needed to conduct further processing on this data, again using PCA, to reduce the clusters to two dimensions.

When we looked at which papers fell in each cluster, however, we had a hard time interpreting these clusters. We could not clearly determine which topic/s in AI ethics were key for each cluster. This was likely due to the high dimensionality, and the small number of papers included in our analysis. We are reminded that contemporary AI relies on *big* data, and thus a larger dataset may be necessary to yield interpretable results with this analysis method. We therefore tried an alternative method for grouping the papers in our canon.

### 4.5. Using AI: LDA Topic Analysis

We next used LDA method to examine the canon, to see if the paper groupings produced made more sense to us. Like *k*-means clustering, LDA is an unsupervised machine learning approach.[18] However, unlike *k*-means, we can use LDA to gather papers under topics, and to then produce a list of words for each topic, making it more interpretable.

LDA is a soft clustering method, which models probability distributions over words and documents. When we use LDA to analyse papers, it treats each paper as an unstructured "bag of words," that is, it does not consider the position of each word in the paper (unlike SciBERT). LDA builds a model of the whole corpus, producing a conditional joint probability distribution of a topic given a word, or a topic given a collection of words (that is, a paper). This means that LDA tries to identify distinct topics by finding correlations between words. Fre-

[17] While this result is not guaranteed by the clustering algorithm, it provided additional reassurance that the groupings reflected real semantic structure in the data.

[18] J.K. Pritchard, M. Stephens, P. Donnelly, *Inference of Population Structure Using Multilocus Genotype Data*, "Genetics" 2000, Vol. 155, No. 2, pp. 945–959, https://doi.org/10.1093/genetics/155.2.945; D. Falush, M. Stephens, J.K. Pritchard, *Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies*, "Genetics" 2003, Vol. 164, No. 4, pp. 1567– 1587, https://doi.org/10.1093/genetics/164.4.1567; D.M. Blei, A.Y. Ng, M.I. Jordan, *Latent Dirichlet Allocation*, "Journal of Machine Learning Research" 2003, Vol. 3, Nos. 4–5, pp. 993–1022, https://doi.org/10.1162/jmlr.2003.3.4-5.993.

quent co-occurrence of words suggests they are related in a topic, whereas non-co-occurrence of words suggests they are not related in a topic.[19]

Our output from LDA is a series of probabilities. For each paper (collection of words) we get a probability that it falls in each topic (here, six possible topics). A paper is therefore not just assigned to one topic – instead, it can have a high probability of concerning multiple topics. This may be for good reason – for example, an overview paper might end up having a high probability of concerning, for instance, "privacy," "AI design" and "robot agency," etc. From examining the topics uncovered in this manner, we felt like we could make some sense of them. We identified the broad themes of each topic as follows:

Topic clusters:

0. Social, social media, gender, culture
1. Superintelligence
2. Applied issues, such as sustainability, health, and the arts
3. Robots, personhood, and artificial agency
4. Design, responsibility
5. Privacy and risk

To prepare the corpus for topic modelling, the cleaned AI ethics texts were first transformed into a document-term matrix using a bag-of-words approach. This matrix represents each document as a vector of word counts, capturing the frequency of the 1,000 most common words across the entire corpus (lower frequency words were not included for reasons of computational tractability).

We then trained the LDA model, specifying that it should extract six topics from the corpus. This decision was informed by the earlier steps in our analysis. In particular, when applying PCA followed by *k*-means clustering, we observed signs of natural groupings in the data. Experimentation with different values of *k*, combined with inspection of silhouette scores, suggested that a range of five to eight clusters produced reasonably coherent and interpretable partitions without over-fragmenting the data. Selecting six topics allowed us to strike a balance

[19] LDA builds a Bayesian probabilistic model of a corpus. It assumes that each document is a mixture of latent topics, and that each topic is characterized by a distribution over words. Formally, LDA posits the following generative process: for each document, a distribution over topics is drawn from a Dirichlet prior; then, for each word in the document, a topic is sampled from that distribution, and a word is sampled from the corresponding topic's word distribution (also drawn from a different Dirichlet prior). The model infers the topic and word distributions that best explain the observed word co-occurrence patterns in the corpus.

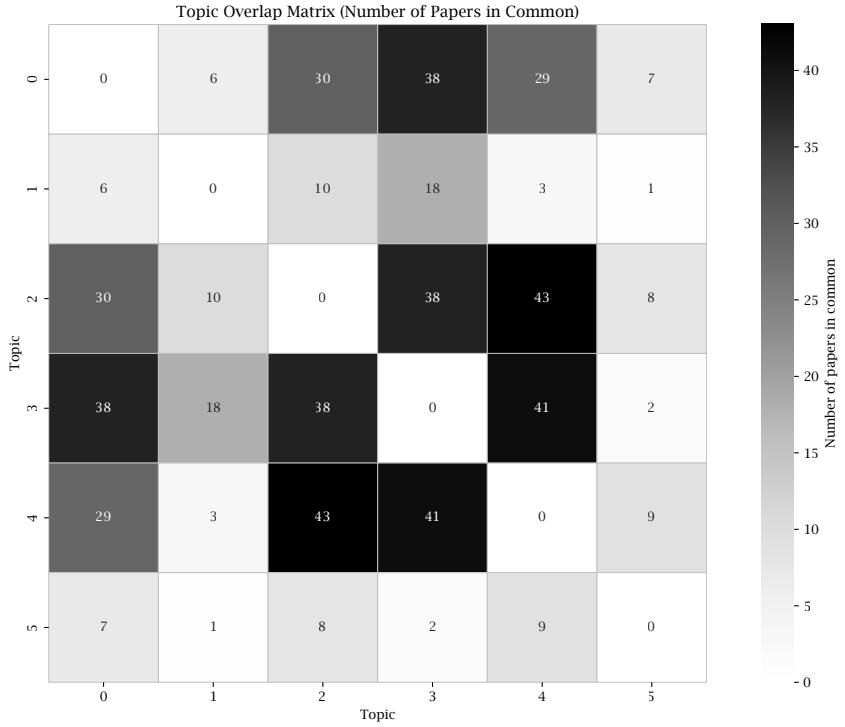



Figure 5. Topic overlaps, showing the number of papers which fell in the overlap of each of the six identified themes

between granularity and conceptual clarity. After fitting the LDA model, each document was assigned a probability distribution over the six topics. To interpret the model, we identified each document's most probable topic – that is, the topic to which it had the highest posterior probability of belonging. This provided a way of associating each paper with a dominant thematic group, based on its characteristic patterns of word usage. We also identified the number of papers in common between topics (Fig. 5).

These topics certainly seemed to us to have some internal unity (as indicated), but they could also be seen not to overlap one another in problematic ways. Looking at the percentage of the papers in one topic (the row in the above table) that overlapped with papers in the other topic (in the columns), we found both that the overlap was not in general too great, and that the overlaps present could also be readily interpreted. For example, 52.9% (18) of the papers on superintelligence

(topic 1) could also be viewed as concerned with a topic involving the notion of artificial agency (topic 3), which is understandable given that ethical concerns around the former appeal to the latter; moreover, looking at the column corresponding to superintelligence, we see that it is entirely white, meaning that none of the other topics overlapped much with it – and indeed, our impression from working within the field is that this topic does, as a matter of sociological fact about the AI ethics community, stand somewhat apart.

## 5. Discussion

### 5.1. Reflection and Learning

The *k*-means clustering, while methodologically sound and internally coherent (as shown by cosine similarity to cluster centroids), ultimately proved difficult to interpret. Although the algorithm grouped texts into clusters based on semantic similarity, we found that the resulting groupings did not consistently align with recognizable course topics or thematic divisions. This may reflect the relatively small size of our corpus, the high dimensionality of the vector space, or the fact that many papers engage with multiple overlapping concerns, making clear separation into exclusive clusters difficult. While the exercise corroborated our preprocessing and embedding pipeline, it may suggest certain limits of hard clustering techniques in the context of philosophical and interdisciplinary content. With this said, it may be that this technique may work more effectively with larger or more varied datasets, or that other dimensional reduction techniques might be needed that better capture the salient structural features of the data, before clustering is applied.

In contrast, the topic modelling using LDA proved to be much more informative. The topics inferred by the model corresponded to intuitively meaningful groupings, such as privacy and risk, robot personhood, or design and responsibility. This method exposed thematic threads that cut across the weekly course topics. Importantly, because LDA provides probabilistic topic distributions, it allowed us to see how individual papers often straddled multiple themes, capturing relations that course structures may obscure. In this sense, LDA may be especially well suited to philosophical corpora, where overlapping normative, conceptual, and technical concerns are the norm rather than the exception.

Of course, it is important to recognize that no computational tools are methodologically neutral: their meaningful interpretation rests upon assumptions about how the data is structured and what counts as significant. For example, TF–IDF and LDA both treat terms as discrete lexical units, abstracted from their syntactic and argumentative context. On the other hand, SciBERT vectorization is sensitive to local linguistic context but will inevitably encode biases from its architecture and training data. We treated semantic similarity as a linearly decomposable property, geometrically represented by cosine similarity in a high-dimensional vector space. This considers meanings as comparable via vector directions and distances, implying that semantic relationships, such as the distinction between "privacy" and "transparency," can be consistently represented as angular differences across the embedding space. In this sense, even our transformer methods may be insensitive to some contextual subtleties.[20] Likewise, *k*-means clustering imposes a fixed number of discrete non-overlapping, roughly isotropic clusters, an assumption unlikely to hold in domains with overlapping, intersecting or multifarious concerns. PCA assumes that the most meaningful structure in the data lies along orthogonal axes of maximal variance, treating key concepts as essentially uncorrelated.[21] The principal component dimensions will not necessarily correspond to conceptual or pedagogical importance. Such assumptions may be justified as reasonable approximations of the real data or by the practical utility of the methods. However, it is essential to recognize them when drawing conclusions from the results. We contend that these tools are best understood not as offering definitive answers, but as producing artefacts that require philosophical interpretation.

In terms of the pedagogical utility of the approach we have undertaken, we have found that the process has yielded discussion and reflection of our core modules in AI ethics. For future iterations of our courses, we can utilize topic words to help identify new literature in areas that are directly related to our course topics, which may help to diversify our recommendations for students. Particularly notable are the areas of overlap, which could be emphasized in our courses to enhance student understanding of the AI ethics landscape. The areas where there

[20] K. Ethayarajh, *How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings*, arXiv:1909.00512, https://doi.org/10.48550/arXiv.1909.00512.

[21] I.T. Jolliffe, *Principal Component Analysis*, op. cit.

is little overlap also interestingly suggests that there may be areas of AI ethics which remain distinct from one another, highlighting potential areas for further exploration (though analysis of a larger corpus would be needed to verify this). In addition, we plan to discuss the results of our analysis with our students, reflecting on the overlapping themes of the courses that go beyond the delineated weeks of the course. For example, the six clear topics we uncovered through LDA did not exactly correspond to our seventeen course topics; some were unsurprising (such as agency, personhood and robot rights); however, others (such as design and responsibility) fall under different sections of the course. Such insights (for example how responsibility can hinge on design choices) may provide stimulating discussion on our courses. Given their aims, noted above, of simultaneously providing the philosophical and computational education needed for students to engage with the realities of responsible AI, we also expect that it will be valuable to discuss the methodological issues we have encountered along the way – such as the difficulties of using *k*-means clustering on sparse data distributed in a high-dimensional space. We may also discuss with them the value of AI assistance, as opposed to full automation, as regards our own ongoing course design: for example, we in no way regard the identification, within our data, of fewer topics than were initially conceived by our course leaders as in any way impugning the expert human judgement that went into our course design; rather, we plan to use the AI-generated insights discussed above to supplement our own decision-making in adapting and revising our syllabi in the ways indicated. This, of course, is a point that applies much more broadly, both within applications of AI for philosophical education, and indeed in other domains more generally.

### 5.2. Computational Analysis for Philosophy

Computing and philosophy have long been intertwined.[22] There are professional bodies dedicated to (aspects of) their intersection, such as the International Association of Computing and Philosophy, as well as the Society for the Philosophy of Artificial Intelligence.[23] And there are, of course, some notable examples of ex-

[22] As a matter of fact, in our own university, the two disciplines initially sat within the same academic unit, or faculty.

[23] International Association of Computing and Philosophy (IACAP), URL: https://www.iacap.org/; Society for the Philosophy of Artificial Intelligence (PHAI), URL: https://philai.net/.

cellent – and early – digital resources in philosophy:[24] the *Stanford Encyclopedia of Philosophy* (*SEP*, created by Edward N. Zalta and Uri Nodelman in 1995); the online (and open access) journal "Philosophers' Imprint" (established in 2001); and *PhilPapers* (begun in 2009).[25] Nevertheless, relatively few philosophers have followed the famous suggestion from Gottfried Wilhelm Leibniz:

> If controversies were to arise, there would be no more need of disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down with their slates and say to each other […]: Let us calculate![26]

That is, "philosophers have arguably failed to take full advantage of the opportunities afforded" by the computational methods that are both available and widely used in the (other) humanities (disciplines).[27] For example, in one list of 145 academic journals dedicated to the digital humanities, a search for "philosophy" yields 0 entries (whereas "humanities" gets 15 hits, "history" has 4, and "literature" 2).[28] Nor are there many pertinent results on Google Scholar when one searches for "digital humanities philosophy," "digital philosophy" or even "computational philosophy." This last term has, however, gained some fluency, and there is even an *SEP* article dedicated to the topic:[29] though that piece is

[24] As noted in J. Weinberg, *Digital Humanities in Philosophy: What's Helpful and What's Hype?*, "Daily Nous", 24.05.2016, URL: https://dailynous.com/2016/05/24/digital-humanities-in-philosophy whats-helpful-whats-hype/.

[25] *The Stanford Encyclopedia of Philosophy*, URL: https://plato.stanford.edu/; Philosophers' Imprint, URL: https://journals.publishing.umich.edu/phimp/; URL: PhilPapers https://philpapers.org/.

[26] Translation cited after B. Russell, *A Critical Exposition of the Philosophy of Leibniz*, Cambridge University Press, Cambridge 1900, pp. 169–170.

[27] B. Ball et al., *Computational Philosophy: Reflections on the PolyGraphs Project*, "Humanities and Social Science Communications" 2024, Vol. 11, No. 186, https://doi.org/10.1057/s41599-024-02619-z, p. 2.

[28] Available at: *The List of Digital Humanities Journals*, URL: https://dhjournals.github.io/list/ (see G. Spinaci, G. Colavizza, S. Peroni, *Preliminary Results on Mapping Digital Humanities Research*, in: *Proceedings of L'Associazione per l'Informatica Umanistica e La Cultura Digitale*, 2020, pp. 246–252, URL: https://aiucd2020.unicatt.it/aiucd-Spinaci_et_al.pdf; G. Spinaci, G. Colavizza, S. Peroni, *A Map of Digital Humanities Research across Bibliographic Data Sources*, "Digital Scholarship in the Humanities" 2022, Vol. 37, No. 4, pp. 1254–1268, https://doi.org/10.1093/llc/fqac016).

[29] P. Grim, D. Singer, *Computational Philosophy*, in: *The Stanford Encyclopedia of Philosophy* (Summer 2024), eds. E.N. Zalta, U. Nodelman, URL: https://plato.stanford.edu/archives/sum2024/entries/computational-philosophy/.

largely concerned with (what has been dubbed) "simulation as a core philosophical method";[30] a quick search of its contents reveals *no* mentions of "natural language processing" (NLP) or "large language models" (LLMs) – techniques and tools that are very widely used in the digital humanities, for both research and teaching purposes… and of course in the pedagogical research we have embarked upon here.

Still, there are some existing digital projects in philosophy, and we shall accordingly devote some (brief) space to their discussion. Many involve data visualizations – for example, the *Philosopher's Web* is a (self styled) "comprehensive map of all influential relationships in philosophy according to Wikipedia."[31] In brief, it shows key figures in philosophy, providing short bios (for some of them), and showing connections (specifically, relations of influence) between them.[32] It does not have a pedagogical focus, but could nevertheless be useful for teaching (perhaps especially the history of) philosophy. Many also involve *SEP* data.[33] Thus, *Visualizing SEP* does precisely what its name says it will:[34] *Stanford Encyclopedia* articles are classified (based on the taxonomy developed by the *Internet Philosophy Ontology Project*[35]), and links to other articles on the same topic(s) are shown. This might be pedagogically useful for students (or researchers) engaged in a literature search – that is, for those trying to figure out what to read

[30] C. Mayo-Wilson, K.J.S. Zollman, *The Computational Philosophy: Simulation as a Core Philosophical Method*, "Synthese" 2021, Vol. 199, pp. 3647–3673, https://doi.org/10.1007/s11229-020-02950-3.

[31] *Philosopher's Web*, URL: https://kumu.io/GOliveira/philosophers-web#map-b9Ts7W5r.

[32] It is described in more detail in J. Jones, *The Philosopher's Web, an Interactive Data Visualization Shows the Web of Influences Connecting Ancient & Modern Philosophers*, Open Culture, 20.10.2017, URL: https://www.openculture.com/2017/10/the-philosophers-web.html; J. Weinberg, *A Visualization of Influence in the History of Philosophy*, "Daily Nous," 11.01.2017, URL: https://dailynous.com/2017/01/11/visualization-influence-history-philosophy/.

[33] As in *The Directed Graph of SEP Related-Entries* (URL: https://mboudour.github.io/2020/05/06/Graph-of-references-among-entries-of-the-Stanford-Encyclopedia-of-Philosophy.html), which provides a (somewhat difficult to see) network representation of the articles in the *SEP*, and the links between them. The dataset underlying this visualization is no doubt of interest – but we prefer to discuss the alternative example in the main text.

[34] *Visualising SEP*, URL: https://www.visualizingsep.com/#.

[35] *Internet Philosophy Ontology Project*, URL: https://www.inphoproject.org/taxonomy. This is a research project funded by the National Endowment for the Humanities. It is committed to open data – so its code is available, as are the taxonomies generated; and there are research papers on the site describing the approach taken. The project is not primarily pedagogical in focus, and only students with fairly advanced technical skills would be well placed to engage with it in any detail.

as they begin on a new topic. (Indeed, philosophy teachers might conceivably look to it when constructing a new course.) Finally, *History of Philosophy: Summarized and Visualized* is a hand-curated visualization of the positions held by philosophers, and their connections – both supporting and conflicting – with theses espoused by other philosophers.[36] Again, it is not primarily a pedagogical project, but might well have pedagogical uses: in particular, it is potentially useful to students to have the substance of the various philosophers' views articulated, and the semantic, or logical, relations between them displayed (and navigable).[37]

Some projects, like ours, are research oriented, and even involve NLP. For example, Mark Alfano has called for collaborators to engage in a semantic mapping project in philosophy, using texts available from Project Gutenberg – and promises to create freely shareable teaching materials![38] The digital humanities approach underlying the project is described in another blog post: it is not unlike the Word2Vec description given in the main text above, though it relies, perhaps, on a different computational technique.[39]

Amongst projects with a pedagogical focus, some are relatively straightforward: *TeachPhilosophy101*, for example, is principally a website with materials – including digital resources – that may be useful to teachers of philosophy.[40] Others involve more comprehensive data analysis: for example, *Open Syllabus Galaxy* maps the most assigned readings across over 7 million course syllabuses.[41] And still others are more targeted: for example, *ArguMap* is a pedagogical app concerned quite specifically with argument mapping;[42] and *The Logic Calculator*

36 D.C. Önduygu, *New Force-Directed Graph with Philosophers as Nodes*, "Deniz Cem Önduygu," 29.01.2025, URL: https://www.denizcemonduygu.com/philo/new force-directed-graph-with-philosophers-as-nodes/.

37 Other projects in the same spirit as those discussed in this paragraph are touched upon in J. Weinberg, *Graphing the History of Philosophical Influences*, "Daily Nous," 21.04.2014, URL: https://dailynous.com/2014/04/21/graphing-the-history-of-philosophical-influences/.

38 M. Alfano, *Collaborators Sought for Digital Humanities Project on the History of Philosophy*, "Philosophy and Other Thoughts," 23.06.2018, URL: https://www.alfanophilosophy.com/blog/2018/6/23/collaborators-sought-for-digital-humanities project-on-the-history-of-philosophy.

39 M. Alfano, *A Semantic-Network Approach to the History of Philosophy, or, What Does Nietzsche Talk about When He Talks about Emotion?*, "Daily Nous," 26.07.2017, URL: https://dailynous.com/2017/07/26/semantic-network-approach-history-philosophy-guest-post mark-alfano/.

40 See *TeachPhilosophy101*, URL: https://www.teachphilosophy101.org/.

41 See *Open Syllabus Galaxy*, URL: https://galaxy.opensyllabus.org/.

42 C. Mohler, *From Maps to Apps: Introducing Students to Argument-Mapping in the Physical and Digital Realms*, "Daily Nous," 25.11.2020, URL: https://dailynous.com/2020/11/25/maps-apps--introducing-students-argument-mapping-guest-post/; *ArguMap*, URL: https://appsolutelyfun.com/argumap.html.

tests for syntactic well-formedness and semantic validity in the propositional calculus.[43] And some seem mostly designed for fun. For example, Justin Weinberg highlights Maximilian Noichl's *SEP* haiku project.[44] This project involves searching the *SEP* for strings of 17 syllables and then checking whether the word breaks fall in the right places to make a haiku. If so, it makes that haiku. The materials produced could be used by teachers looking to find appropriate tidbits to introduce lectures, or to serve as mnemonics for students.

This is by no means an exhaustive overview of the digital projects that have been pursued in relation to philosophy, or philosophical pedagogy, but it is not entirely unrepresentative in our view. And if we are right about that, it should be clear that what we have done here is quite atypical, at least within philosophy. For we have turned quite heavy-duty computational methods upon our own teaching practice – the syllabi we have created – to see what they reveal about the contents of our courses.

Indeed, we pause to briefly dwell on the novelty of the approach taken here, not only relative to existing practices within philosophy, but even in the context of digital humanities as a whole. Advances in AI have come fast and thick in recent years, bringing disruption across all aspects of society. Digital humanities can hardly be expected to prove an exception – and indeed, some scholars have begun to grapple with the question of how to incorporate advanced NLP techniques into humanities research.[45] And yet, to the best of our knowledge, ours is the first attempt within the humanities to use transformer-based vector embeddings of whole documents to provide distant readings for the analysis of a corpus.[46] While this particular method has not yielded deep insights in looking

[43] I. Votsis, *The Logic Calculator*, 2019, URL: https://votsis.org/logic.html.

[44] J. Weinberg, *Making Haiku and Art from the SEP*, "Daily Nous," 31.08.2021, URL: https://dailynous.com/2021/08/31/making-haiku-art-sep/.

[45] O. Suissa, A. Elmalech, M. Zhitomirsky-Geffet, *Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science*, "Journal of the Association for Information Science and Technology" 2022, Vol. 73, No. 2, pp. 268–287, https://doi.org/10.1002/asi.24544; A. Ehrmanntraut et al., *Type-and Token-Based Word Embeddings in the Digital Humanities*, in: *CHR 2021: Computational Humanities Research Conference*, 2021, pp. 16–38, URL: https://ceur-ws.org/Vol-2989/long_paper35.pdf; C. Liu et al., *SikuGPT: A Generative Pre-Trained Model for Intelligent Information Processing of Ancient Texts from the Perspective of Digital Humanities*, "ACM Journal on Computing and Cultural Heritage" 2024, Vol. 17, No. 4, pp. 1–17, https://doi.org/10.1145/3676969.

[46] However, see efforts by Arman Cohan et al. to adapt similar methods in other fields: *SPECTER: Document-Level Representation Learning Using Citation-Informed Transformers*, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds. D. Jurafsky et al., Association for Computational Linguistics, 2020, pp. 2270–2282, https://doi.org/10.18653/v1/2020.acl-main.207.

at our relatively modestly sized corpus – and certainly none that can themselves be generalized to, for example, other (individual) syllabus analyses – we anticipate that this pioneering approach, as we develop and refine it further, or at least its ultimate assessment (for example, through comparison with the older LDA-based technique), will prove valuable well beyond the present context.

In future research we will continue exploring the possibilities of this analytic approach for AI ethics and philosophical pedagogy. In particular, we plan to analyse a wider corpus of texts in relevant fields. We hope that this will help us gain a better understanding of relevant literature, identify emerging topics as well as literature gaps, and draw on uncovered connections between topics and bodies of work to signpost to our students.

## 6. Conclusion

We have demonstrated how computational analysis of readings on philosophical syllabi can yield useful reflections for educators in philosophy. Our dataset consisted of the materials assigned in two of our philosophy courses in the field of AI ethics. We prepared this dataset for analysis, taking into account any ethical concerns with our proposed approach. We implemented several NLP techniques to analyse our corpus. We began with relatively simple approaches (word frequency analysis and TF–IDF) which yielded some noteworthy results, particularly the relative importance of the "human" in the AI ethics course corpus, and the relative unimportance (compared to the Wittgenstein corpus) of "philosophy." Given the nature of these approaches, only limited conclusions could be drawn. We then moved on to more complex NLP approaches, including document vectorization via SciBERT, clustering via *k*-means, and topic modelling using LDA. SciBERT vectorization and clustering allowed us to explore semantic relationships within the corpus; however, we struggled to draw conclusions from this approach, likely due to the small number of papers in our corpus. In future analyses we plan to use a larger dataset in order to combat this limitation. Topic modelling through LDA enabled us to identify six broad themes in the corpus, which were in some cases different to what we might expect given the topics we set and how these are connected on the course. Finally, we discussed the broader implications of our approach, both for AI ethics education and for philosophy as a discipline. Given

the limits of existing work in computational approaches in the field of philosophical research (even in AI ethics), we see an opportunity to harness these approaches for philosophy and philosophical education.

**Acknowledgements**
We are grateful for the support of this project provided by Northeastern University's Ethics Institute, the Internet Democracy Initiative and NULab. We would also like to thank the participants at the NULab Spring Conference, and the audience of the webinar of the "Philosophical Education" journal, for their helpful feedback.

## Bibliography

Alfano M., *Collaborators Sought for Digital Humanities Project on the History of Philosophy*, "Philosophy and Other Thoughts," 23.06.2018, URL: https://www.alfanophilosophy.com/blog/2018/6/23/collaborators-sought-for-digital-humanities project-on-the-history-of-philosophy.

Alfano M., *A Semantic-Network Approach to the History of Philosophy, or, What Does Nietzsche Talk about When He Talks about Emotion?*, "Daily Nous," 26.07.2017, URL: https://dailynous.com/2017/07/26/semantic-network-approach-history-philosophy-guest-post mark-alfano/.

Ball B., Helliwell A.C., Rossi A., *Wittgenstein and Artificial Intelligence: Mind and Language*, Anthem Press, London 2024.

Ball B., Helliwell A.C., Rossi A., *Wittgenstein and Artificial Intelligence: Values and Governance*, Anthem Press, London 2024.

Ball B., Koliousis A., Mohanan A., Peacy M., *Computational Philosophy: Reflections on the PolyGraphs Project*, "Humanities and Social Science Communications" 2024, Vol. 11, No. 186, https://doi.org/10.1057/s41599-024-02619-z.

Beltagy I., Lo K., Cohan A., *SciBERT: A Pretrained Language Model for Scientific Text*, arXiv:1903.10676, https://doi.org/10.48550/arXiv.1903.10676.

Bengio Y., Ducharme R., Vincent P., Jauvin C., *A Neural Probabilistic Language Model*, "Journal of Machine Learning Research" 2003, Vol. 3, pp. 1137–1155.

Blei D.M., Ng A.Y., Jordan M.I., *Latent Dirichlet Allocation*, "Journal of Machine Learning Research" 2003, Vol. 3, Nos. 4–5, pp. 993–1022, https://doi.org/10.1162/jmlr.2003.3.4-5.993.

Cohan A., Feldman S., Beltagy I., Downey D., Weld D., *SPECTER: Document-Level Representation Learning Using Citation-Informed Transformers*, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds. D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Association for Computational Linguistics, 2020, pp. 2270–2282, https://doi.org/10.18653/v1/2020.acl-main.207.

Devlin J., Chang M.-W., Lee K., Toutanova K., *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805, https://doi.org/10.48550/arXiv.1810.04805.

Ehrmanntraut A., Hagen T., Konle L., Jannidis F., *Type-and Token-Based Word Embeddings in the Digital Humanities*, in *CHR 2021: Computational Humanities Research Conference*, 2021, pp. 16–38, URL: https://ceur-ws.org/Vol-2989/long_paper35.pdf.

Ethayarajh K., *How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings*, arXiv:1909.00512, https://doi.org/10.48550/arXiv.1909.00512.

Falush D., Stephens M., Pritchard J.K., *Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies*, "Genetics" 2003, Vol. 164, No. 4, pp. 1567– 1587, https://doi.org/10.1093/genetics/164.4.1567.

Firth J.R., *Chapter 11: A Synopsis of Linguistic Theory*, in: *Selected Papers of J.R. Firth, 1952–59*, ed. F.R. Palmer, Longmans, London 1968, pp. 168–205.

Firth J.R., *A Synopsis of Linguistic Theory 1930–1955*, in: *Studies in Linguistic Analysis*, Blackwell, Oxford 1957, pp. 1–32.

Grim P., Singer D., *Computational Philosophy*, in: *The Stanford Encyclopedia of Philosophy* (Summer 2024), eds. E.N. Zalta, U. Nodelman, URL: https://plato.stanford.edu/archives/sum2024/entries/computational-philosophy/.

Jolliffe I.T., *Principal Component Analysis*, 2nd ed., Springer, New York 2002.

Jones J., *The Philosophers Web, an Interactive Data Visualization Shows the Web of Influences Connecting Ancient & Modern Philosophers*, Open Culture, 20.10.2017, URL: https://www.openculture.com/2017/10/the-philosophers-web.html.

Liu C., Wang D., Zhao Z., Hu D., Wu M., Lin L., Liu J., Zhang H., Shen S., Li B., Zhao L., *SikuGPT: A Generative Pre-Trained Model for Intelligent Information Processing of Ancient Texts from the Perspective of Digital Humanities*, "ACM Journal on Computing and Cultural Heritage" 2024, Vol. 17, No. 4, pp. 1–17, https://doi.org/10.1145/3676969.

MacQueen J.B., *Some Methods for Classification and Analysis of Multivariate Observations*, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1: *Statistics*, University of California Press, Berkeley–Los Angeles 1967, pp. 281–297.

Mayo-Wilson C., Zollman K.J.S., *The Computational Philosophy: Simulation as a Core Philosophical Method*, "Synthese" 2021, Vol. 199, pp. 3647–3673, https://doi.org/10.1007/s11229-020-02950-3.

Mikolov T., Chen K., Corrado G., Dean J., *Efficient Estimation of Word Representations in Vector Space*, arXiv:1301.3781, https://doi.org/10.48550/arXiv.1301.3781.

Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J., *Distributed Representations of Words and Phrases and Their Compositionality*, arXiv:1310.4546, https://doi.org/10.48550/arXiv.1310.4546.

Mohler C., *From Maps to Apps: Introducing Students to Argument-Mapping in the Physical and Digital Realms*, "Daily Nous," 25.11.2020, URL: https://dailynous.com/2020/11/25/maps-apps-introducing-students-argument-mapping-guest-post/.

Önduygu D.C., *New Force-Directed Graph with Philosophers as Nodes*, "Deniz Cem Önduygu," 29.01.2025, URL: https://www.denizcemonduygu.com/philo/new force-directed-graph-with-philosophers-as-nodes/.

Pedregosa F., et al., *Scikit-Learn: Machine Learning in Python*, "Journal of Machine Learning Research" 2011, Vol. 12, pp. 2825–2830.

Pritchard J.K., Stephens M., Donnelly P., *Inference of Population Structure Using Multilocus Genotype Data*, "Genetics" 2000, Vol. 155, No. 2, pp. 945–959, https://doi.org/10.1093/genetics/155.2.945.

Rousseeuw P.J., *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*, "Computational and Applied Mathematics" 1987, Vol. 20, pp. 53–65, https://doi.org/10.1016/0377-0427(87)90125-7.

Russell B., *A Critical Exposition of the Philosophy of Leibniz*, Cambridge University Press, Cambridge 1900.

Spärck Jones K., *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, "Journal of Documentation" 1972, Vol. 28, No. 1, pp. 11–21, https://doi.org/10.1108/eb026526.

Spinaci G., Colavizza G., Peroni S., *A Map of Digital Humanities Research across Bibliographic Data Sources*, "Digital Scholarship in the Humanities" 2022, Vol. 37, No. 4, pp. 1254–1268, https://doi.org/10.1093/llc/fqac016.

Spinaci G., Colavizza G., Peroni S., *Preliminary Results on Mapping Digital Humanities Research*, in: *Proceedings of L'Associazione per l'Informatica Umanistica e La Cultura Digitale*, 2020, pp. 246–252, URL: https://aiucd2020.unicatt.it/aiucd-Spinaci_et_al.pdf.

Steinhaus H., *Sur la division des corps matériels en parties*, "Bulletin de l'Académie Polonaise des Sciences, Classe III" 1956, Vol. 4(12), pp. 801–804.

Suissa O., Elmalech A., Zhitomirsky-Geffet M., *Text Analysis Using Deep Neural Networks in Digital Humanities and Information Science*, "Journal of the Association for Information Science and Technology" 2022, Vol. 73, No. 2, pp. 268–287, https://doi.org/10.1002/asi.24544.

Weinberg J., *Digital Humanities in Philosophy: What's Helpful and What's Hype?*, "Daily Nous," 24.05.2016, URL: https://dailynous.com/2016/05/24/digital-humanities-in-philosophy whats-helpful-whats-hype/.

Weinberg J., *Graphing the History of Philosophical Influences*, "Daily Nous," 21.04.2014, URL: https://dailynous.com/2014/04/21/graphing-the-history-of-philosophical-influences/.

Weinberg J., *Making Haiku and Art from the SEP*, "Daily Nous," 31.08.2021, URL: https://dailynous.com/2021/08/31/making-haiku-art-sep/.

Weinberg J., *A Visualization of Influence in the History of Philosophy*, "Daily Nous," 11.01.2017, URL: https://dailynous.com/2017/01/11/visualization-influence-history-philosophy/.

Votsis I., *The Logic Calculator*, 2019, URL: https://votsis.org/logic.html.

Edukacja Filozoficzna
ISSN 0860-3839, eISSN 2956-8269
DOI: 10.14394/edufil.2025.0004
ORCID: 0009-0001-8886-6620

# Justice and AI Fairness: John Rawls and Iris Marion Young on Racist and Sexist AI Decisions

Neomal Silva
(Independent researcher and Senior ICT Consultant, Melbourne, Australia)

**Abstract:** AI outcomes that exhibit racism, sexism, homophobia, or other biases are deemed "unfair." Several scholars have applied John Rawls's theory of justice to evaluate this unfairness. This paper clarifies, though, that Rawls's ideal and nonideal theories are ill-equipped to deal with individual instances of AI unfairness; it furthermore argues that Young's theory is better equipped to do so – not only because it includes sociological accounts of racism and other -isms, but also because it incorporates the consciousness-raising spaces that help "name" the racist, sexist, etc. behaviours – behaviours that, if left unnamed, remain undetected, and, as a result, are both re-enacted in society and reproduced by AI.

**Key words:** AI bias, AI fairness, Rawls, structural power, Iris Marion Young

## 1. Introduction

Whilst artificial intelligence (AI) encompasses robotics, rule-based systems, machine learning, and other technologies, it is machine learning, in particular, that has provided several instances of bias against marginalized groups – such as non-white people and females. Consider the following practical instances of AI bias.[1]

Amazon's recruitment team used an algorithm to rate CVs from one to five stars, only to find it favoured male candidates. The bias stemmed from training

[1] I use the terms "fairness" and "bias" (or "AI fairness" and "AI bias") interchangeably: "AI fairness" aims to ensure that no group – defined by some socially salient trait like gender or ethnicity – is unfairly disadvantaged. "AI bias," on the other hand, refers to the unfair or skewed outcomes that discriminate against certain groups. In essence, "AI fairness" is the goal, whilst "AI bias" refers to the obstacles to achieving it. For an overview of definitions of different types of AI fairness and AI bias, along with a survey of different data-centric techniques for mitigating bias, see E. Ferrara, *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies*, "Sci" 2024, Vol. 6, No. 1, pp. 1–15, https://doi.org/10.3390/sci6010003.

the algorithm on a dataset made up of CVs of people previously hired for the role – most of whom were men.[2]

Joy Buolamwini discovered that commercial facial recognition technologies from companies like IBM, Microsoft, and Megvii had higher error rates for darker-skinned people and women.[3] The bias arose because the algorithms were primarily trained on faces of young white men.

Google launched a photos app designed to categorize users' photos but faced backlash when it miscategorized African Americans as "gorillas." This offensive error occurred because the algorithm lacked sufficiently diverse training data.

To examine instances of AI *unfairness* such as these, scholars might turn to John Rawls's concept of justice as *fairness*. Whilst some have used Rawls's work to study AI ethics,[4] Morten Bay cautions against oversimplifying or taking Rawls's ideas out of context.[5] Nonetheless, scholars have engaged with Rawls in their studies of AI bias and fairness.[6] For example, Flavia Barsotti and Rüya Gökhan Koçer argue that Rawls's *Theory of Justice* "provides the foundations to a solution

---

2 J. Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*, Reuters, 9.10.2018, URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

3 J. Buolamwini, T. Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, "Proceedings of Machine Learning Research" 2018, Vol. 81, p. 8.

4 E.g., I. Gabriel, *Toward a Theory of Justice for Artificial Intelligence*, "Daedalus" 2022, Vol. 151, No. 2, pp. 218–231, https://doi.org/10.1162/daed_a_01911; H. Heidari et al., *Fairness behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making*, "Advances in Neural Information Processing Systems" 2018, Vol. 31; R. Binns, *Algorithmic Accountability and Public Reason*, "Philosophy and Technology" 2018, Vol. 31, p. 543; L. Weidinger et al., *Using the Veil of Ignorance to Align AI Systems with Principles of Justice*, "Proceedings of the National Academy of Sciences of the United States of America" 2023, Vol. 120, e2213709120, https://doi.org/10.1073/pnas.2213709120.

5 M. Bay, *Participation, Prediction, and Publicity: Avoiding the Pitfalls of Applying Rawlsian Ethics to AI*, "AI and Ethics" 2024, Vol. 4, p. 1545, https://doi.org/10.1007/s43681-023-00341-1.

6 E.g., see F. Barsotti, R.G. Koçer, *MinMax Fairness: From Rawlsian Theory of Justice to Solution for Algorithmic Bias*, "AI & Society" 2024, Vol. 39, pp. 961–974, https://doi.org/10.1007/s00146-022-01577-x; A.K. Jørgensen, A. Søgaard, *Rawlsian AI Fairness Loopholes*, "AI and Ethics" 2022, Vol. 3, pp. 1185–1192, https://doi.org/10.1007/s43681-022-00226-9; T. Krupiy, *A Vulnerability Analysis: Theorising the Impact of Artificial Intelligence Decision-Making Processes on Individuals, Society and Human Diversity from a Social Justice Perspective*, "Computer Law & Security Review" 2020, Vol. 38, 105429, https://doi.org/10.1016/j.clsr.2020.105429; L.M. Rafanelli, *Justice, Injustice, and Artificial Intelligence: Lessons from Political Theory and Philosophy*, "Big Data and Society" 2022, Vol. 9, No. 1, https://doi.org/10.1177/20539517221080676.

for algorithmic bias"[7] – where the algorithmic bias is against "gender, ethnicity, disability, etc."[8] To offer another example: Anna Katrine Jørgensen and Anders Søgaard, though they critique the use of Rawls to achieve algorithmic fairness, assume his difference principle can be applied to "groups […] typically thought of as the product of a subset of protected attributes, e.g., gender and race."[9] However, Rawls's difference principle[10] is concerned with income groups, not groups defined by protected attributes. In Rawls's framework, the "worst off" refers to those with the least income or wealth, and economic inequality is allowed only if it benefits the absolute position of that socioeconomically disadvantaged group. It should be apparent, then, that scholars should tread carefully when applying Rawls's ideas to AI fairness.

One aim of this paper is not only to urge AI fairness scholars to exercise caution when applying Rawlsian concepts, like the difference principle or the veil of ignorance, but also to argue a stronger claim: fundamentally, Rawls's theory is ill-equipped to address biases related to race, gender, and other forms of discrimination in AI. This is partly because Rawls abstracts from structural power – a type of power implicated in racism, sexism, and other -isms[11] – but also because his ideal and nonideal theories are not designed to tackle specific instance of social injustice (like biased machine-learning outputs). Though A. John Simmons[12] has

7 F. Barsotti, R.G. Koçer, *MinMax Fairness*, op. cit., p. 961. It is also too big a jump to go from Rawls's *Theory of Justice* – which concerns the two principles of justice that Rawls argues should govern the basic structure of society (e.g., the constitution) – to immediately proposing that Rawls's two principles ought to constrain the outputs of a machine-learning algorithm. Iason Gabriel, in his *Toward a Theory of Justice for Artificial Intelligence*, points out that technology (and AI, in particular) cannot be assumed to be part of the basic structure – i.e., it cannot be assumed to be the part of the subject of Rawls's two principles of justice – but Gabriel argues strongly for its inclusion. See I. Gabriel, *Toward a Theory of Justice*, op. cit.

8 F. Barsotti, R.G. Koçer, *MinMax Fairness*, op. cit., p. 964.

9 A.K. Jørgensen, A. Søgaard, *Rawlsian AI Fairness Loopholes*, op. cit., p. 1187.

10 J. Rawls, *A Theory of Justice*, Harvard University Press, Cambridge, MA, 1971, p. 83.

11 I define racism and sexism much like Iris Young understands them. She views "racism" as a systemic and structural phenomenon that marginalizes and disadvantages racial groups. This occurs through institutional practices, cultural norms, and social policies that perpetuate racial inequalities. Racism, in this sense, goes beyond overt discrimination or prejudice and includes the ways societal institutions maintain and reproduce racial hierarchies. Similarly, "sexism," in Young's view, is a structural form of oppression that subordinates women and reinforces gender roles through societal norms, institutions, and practices. Not limited to individual acts of discrimination, it furthermore encompasses the pervasive behavioural norms that perpetuate gender inequality and limit women's opportunities.

12 A.J. Simmons, *Ideal and Nonideal Theory*, "Philosophy & Public Affairs" 2010, Vol. 38, pp. 5–36.

argued that Rawls's theories are not suited to addressing specific social injustices outside the context of AI, this critique is yet to be articulated in AI fairness literature. I will articulate it here, as it is vital to prevent scholars from misapplying Rawls's theories to challenges his work is not equipped to solve.

A second aim of this paper is to propose Iris Marion Young's critical theory of social justice as an alternative to Rawls's theory. Unlike Rawls's, Young's theory is deeply connected to sociological accounts of structural power. I will show that engagement with structural power is essential for evaluating unfairness in AI decision-making, making Young's theory the preferable approach. Crucially, her theory provides the conceptual tools to expose the very -isms that are reproduced in the AI outcomes that draw the most media criticism – such as gender-biased recruitment,[13] racist image classification,[14] antisemitic messaging,[15] and over-policing of certain ethnicities.[16]

I proceed as follows. In section 2, I provide Rawls's accounts of what is "just," what is "unjust," and what is "permissible," and I clarify that these accounts are not intended to deal with single instances of unfairness. Notably, none of Rawls's accounts (of what is "just," "unjust," etc.) refer to structural power. In section 3, we consider structural power, using an example to illuminate some of its complexities, along with some of the consequences it can have for those disadvantaged by it. That elucidation helps confirm that Rawls's theory is not equipped to attend to the kinds of injustices that worry AI ethicists. Its disregard for structural power may prompt philosophers to seek a theory that does engage with it. In section 4, we turn to one such theory – Young's feminist critical theory. We note its ability to capture the power that resides at what Anthony Giddens calls the level of "practical consciousness." Moreover, we examine its engagement with discursive consciousness-raising spaces – that is, the spaces in which structural

[13] Reuters, *Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs*, "The Guardian," 11.10.2018, URL: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

[14] M. Zhang, *Google Photos Tags Two African-Americans as Gorillas through Facial Recognition Software*, "Forbes," 1.07.2015, URL: https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.

[15] S. Buranyi, *Rise of the Racist Robots: How AI Is Learning All Our Worst Impulses*, "The Guardian," 8.08.2017, URL: https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

[16] CBC Radio, *Police Are Considering the Ethics of AI, Too*, 21.09.2018, URL: https://www.cbc.ca/radio/spark/tech-in-policing-1.4833189/police-are-considering-the-ethics-of-ai-too-1.4833194.

oppression has historically found its voice. The attributes of Young's critical theory not only enable us to conceptualize structural power but also equip us with the tool – namely, consciousness-raising spaces – that could help liberate society from its various -isms. In section 5, we investigate the implications of these insights for AI decision outcomes. We also address potential objections. Section 6 offers concluding remarks.

## 2. Rawls's Ideal and Nonideal Theories of Justice

Rawls's theory of "justice as fairness" argues that a just society (i) institutes socioeconomic inequalities only if they benefit the "worst off," and (ii) ensures all members have equal basic liberties and fair equality of opportunity. This is achieved through abstractions like the "original position"[17] and "veil of ignorance,"[18] where rational individuals would choose these two principles of justice when they are unaware of their own social status or of their own personal characteristics and talents.

Some philosophers argue that Rawls's theory is ill-equipped to address issues like sexism, racism, and other -isms. It "abstracts from the determinate content of social life,"[19] they say, ignores "the importance of social groups,"[20] and is mute on how "to rectify [racial] injustices that have already occurred."[21] Rawls, of course, offers us an ideal theory of justice (as outlined above) – but also a nonideal theory. His ideal theory elucidates an abstract conception of justice, whilst his nonideal theory articulates how to move us closer to it – without that nonideal theory necessarily attempting to eliminate particular instances of injustice, such as -isms. This requires some explanation.

In his ideal theory, Rawls articulates the constitutional principles that citizens would choose from behind a veil of ignorance, that is, choose under conditions where potential biases influencing their judgment are hidden from view. Rawls

[17] J. Rawls, *A Theory of Justice*, op. cit., pp. 118–194.

[18] Ibid., pp. 136–141.

[19] L. McNay, *Recognition as Fact and Norm: The Method of Critique*, in: *Political Theory: Methods and Approaches*, eds. D. Leopold, M. Stears, Oxford University Press, Oxford 2008, p. 87.

[20] I.M. Young, *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990, p. 27.

[21] See C.W. Mills, *Retrieving Rawls for Racial Justice? A Critique of Tommie Shelby*, "Critical Philosophy of Race" 2013, Vol. 1, No. 1, p. 2 (italics removed).

says that the principles so derived are "just" and that they represent the ideal to which a society ought to strive if it is to be said to be a "just" society.

Simmons, in his essay *Ideal and Nonideal Theory*,[22] responds to the complaint that Rawls's nonideal theory is silent on real-world problems, such as historical slavery,[23] and resource scarcity[24] – and his response is: it's not meant to speak to such problems. Rawls's nonideal theory does not concern itself with removing single instances of injustice per se – where such instances might include crime, or an -ism. Its purpose, instead, is to do what is required to move society from less-than-just to (Rawls's notion of) "just," as long as the actions that are taken to carry out that move are "morally permissible," "politically feasible," and likely to succeed.[25] Simmons acknowledges that Rawls is vague on those three conditions.[26] What matters for present purposes, though, is that Rawls's nonideal theory endorses attending to an -ism only if doing so moves us closer to his ideal. Indeed, non-intervention, or even introducing a new -ism, is permissible, if it is thought to be the necessary transitional path for a society to ultimately achieve (Rawls's) "just" state.[27]

We can now say the following about Rawls's framework. A society is "just" if it has fully realized his two principles of justice. It is "unjust" if it hasn't. It is "permissible" to not intervene to address an -ism.

Furthermore, assessments of what is "just," "unjust," or "permissible" can be made without considering structural power, or engaging with discourses about lived experiences of it. I contend that this omission is problematic (at least for our present purposes of considering racist etc. outcomes). I am not alone in contending this.[28] We will consider an example in which structural power is in play – not

22 A.J. Simmons, *Ideal and Nonideal Theory*, op. cit., p. 19.

23 C.W. Mills, *"Ideal Theory" as Ideology*, "Hypatia" 2005, Vol. 20, No. 3, p. 168.

24 C. Farrelly, *Justice in Ideal Theory: A Refutation*, "Political Studies" 2007, Vol. 55, p. 853.

25 J. Rawls, *The Law of Peoples*, Harvard University Press, Cambridge, MA, 1999, p. 89.

26 A.J. Simmons, *Ideal and Nonideal Theory*, op. cit., p. 19.

27 For support for this interpretation of Rawls's view, and an elaboration of it, see ibid., p. 23.

28 See I.M. Young, *Structure as the Subject of Justice*, in: I.M. Young, *Responsibility for Justice*, Oxford University Press, Oxford 2011, https://doi.org/10.1093/acprof:oso/9780195392388.003.0002, where she argues that structural power is the subject of justice, and that *pace* Rawls his basic structure in his conception of the Just ought to factor it in. Also see L. McNay, *Recognition as Fact and Norm*, op. cit., pp. 85–105, where the author offers a critique of the kind of idealized normative reasoning we find in Rawls's theory in the first section, and in the latter part of her paper she challenges Jürgen Habermas to attend more carefully to the effects of structural power in his communicative ethics.

just to highlight the problem of omitting it, but also because the example helps convey what the complex phenomenon of structural power looks like, along with some of the impacts that it has – impacts which, I hope to show, cannot *pace* Rawls be assumed to have nothing to do with an assessment of what is just, unjust, or morally permissible in our existing social arrangements.

## 3. Structural Power: An Illustrative Example

In the days that followed Martin Luther King's assassination, Jane Elliot, a third-grade schoolteacher in a small rural town in Iowa, exasperated by the persistent cycles of racism within America, felt that she needed to help her classroom students understand racism in a more meaningful way. She had spoken to them about discrimination in the past. But now she wanted them to sense the anguish of the racially discriminated Other, to feel their despair, "to walk in […] [their] moccasins", as she put it.[29]

Elliot divided the students into two categories based on their eye colour. She then announced, "Blue-eyed people are better than brown-eyed people. They are cleaner than brown-eyed people. They are more civilized than brown-eyed people. And they are smarter than brown-eyed people."[30] The blue-eyed children, she added, are to receive an extra five minutes to play at lunchtime, whereas brown-eyed children are barred from playing on the playground equipment from hereon in.

Suppose that Ms Elliot furthermore segregates the classroom, confining the brown-eyed children to the back left corner, and only allowing blue-eyed children to sit at the front. In the days and weeks that follow, Suzy, a particularly intelligent (brown-eyed) student never seems to get seen by Ms Elliot when she raises her hand, perhaps because Ms Elliot has grown accustomed to not looking towards that section of the room when she asks a question.[31]

The above example allows us to provide an initial outline of what structural power looks like. To be clear, brown-eyedness (and blue-eyedness) goes beyond mere "colour" here – it's not about a relationship between one's iris and the sur-

[29] PBS Frontline, *A Class Divided*, "CosmoLearning" 1985.

[30] Ibid.

[31] This is analogous to what happened at Amazon when female job applicants (who can be said to have been "putting up their hand for a job opportunity") were screened out by the AI used by Amazon for recruitment purposes. See Reuters, *Amazon Ditched AI Recruiting Tool*, op. cit.

rounding light. Rather, at least in part, brown-eyedness acquires social significance within this classroom context in relation to blue-eyedness – that is, brown-eyedness is *not* blue-eyedness. Each of these social categories emerges as a social construct intricately interwoven with the discourses generated, perpetuated, compounded, and sometimes contested, by the students, and of course, their teacher. There is a dialectical interchange between the social categories and classroom power dynamics themselves: the categories are created by power dynamics (primarily constructed and imposed, as they were, by the teacher herself), and the categories themselves reinforce and exacerbate those power dynamics (by structuring the teacher–student and student–student interactions). The concept of power between social categories, such as "blue-eyedness" and "brown-eyedness," plays a significant role in understanding the advantages or disadvantages that members of those social categories encounter – not just the possibility of using the playground equipment, but, as Suzy finds, the power to be seen, heard, respected, and listened to as an equal.[32]

The classroom with its eye-ism is analogous to actual societies riddled with the structural power of various -isms.[33] Rawls's theory remains unswayed by such power, though. The above situation is "unjust" on his account. However, that is not due to the existence of eye-ism – but to the non-realization of Rawls's two principles of justice. Furthermore, as Simmons argues in his reading of Rawls, it would be "impermissible" to remove eye-ism if that resulted in Raymond rebelling against its removal by rallying his blue-eyed compatriots to beat up the brown-eyes and strip them further of basic liberties.

There are two messages that one can take from this. There are many non-political-theorists and activists[34] who study AI bias, and our first message is for them. Already troubled by racist image classification, sexist CV filtering, etc. – they might now also be exasperated to learn that Rawls's theory would not judge

[32] As Lois McNay points out in her critique of Habermas's ideal speech situation, power dynamics permeate interpersonal exchanges, existing before them and continuing throughout. See L. McNay, *Recognition as Fact and Norm*, op. cit., pp. 85–105.

[33] We will treat the "classroom" as though it is a "state" as we work through our reasoning – since Rawls's theory of justice applies to states (rather than classrooms).

[34] In referring to "activists," I have in mind scholars like Joy Buolamwini (Founder of the Algorithmic Justice League) – who self-identifies as an activist – but also researchers like Timnit Gebru (co-founder of Black in AI), Deborah Raji, and Safiya Noble (who says in her book *Algorithms of Oppression* that she hopes to end social injustice and change the perception of marginalized people in technology).

any of those AI outcomes to be "impermissible" in and of themselves. Our analysis hopefully makes clear that Rawls's theory is ill-suited to realize their aims. His theory is fit-for-purpose if one's purpose is to clarify what (in Rawls's view) the most perfectly just society looks like. However, it is not the correct tool if your task is to eliminate particular injustices (such as those that arise in AI decision-making). The second message is to philosophers, concerned that Rawls's framework ignores structural power if it is called upon to determine the permissibility of AI outcomes. This does not, of course, mean that structural power can be assumed to have an impact on its moral permissibility – only that it perhaps should not be ignored from the outset. For that reason, they may wish to turn to critical theory, which can consider, and critically analyse, power, when it decides on the moral permissibility of AI outcomes.

## 4. Young's Feminist Critical Theory

A critical theoretical approach, such as that of Iris Marion Young, is dialectically linked to sociological analysis. An assessment of the social injustice of an interpersonal arrangement, she maintains, demands a social theory about the structural power within it. Young relies on Anthony Giddens's theory of structuration,[35] as well as Pierre Bourdieu's concept of habitus, to theorize -isms, making normative recommendations on its basis – rather than in the abstract.

A thorough account of her interpretation and fusion of those two social theories can be found in her essay *Structure as the Subject of Justice*.[36] People in a social setting follow certain "rules" of engagement, many of which are implicit, but for which one risks sanction if violated; for example, queue jumping; or not saying "please" when asking a favour. When people's following of such rules is implicit, it can be said to take place at the level of "practical consciousness" – meaning the actor performs the action, without being able unambiguously to explain its logic. Furthermore, within a social setting, a person has what Giddens calls "resources" – understood (at the societal level) as both the material items one relies upon to create and produce physical goods and technologies, and the nonmate-

[35] A. Giddens, *The Constitution of Society: Outline of the Theory of Structuration*, Polity, Cambridge 1986.

[36] I.M. Young, *Structure as the Subject of Justice*, op. cit.

rial social skills that bolster a person's social power (where the latter skills could include gravitas, and the ability to persuade or manipulate others).[37] Those people in a social setting who understand its rules, and possess more resources, can be said to more powerful than those who don't.

Across her body of work, Young seeks to address the concerns of social groups within contemporary American society.[38] A "social group" is not a mere collection of individuals. It is a socially salient category that structures relations between "those to whom the category attaches" and "other people within the social setting" – relations that can be described in terms such as discrimination, stereotyping, stigmatization, exclusion, socioeconomic disadvantage, and other forms of disadvantage. The social groups that focus Young's critical theory of contemporary American society include "Blacks, Latinos, American Indians, poor people, lesbians, old people, […] the disabled"[39] and, of course, women. Throughout her *Justice and the Politics of Difference*, Young argues that such citizens tend to possess fewer resources, and find themselves in social settings in which they are less adept at following the settings' rules than the dominant group. In other words, they are less powerful due to their social group membership. I do not find this claim controversial. There are many examples of such power differentials, including those that tie to perceived rule violation by the Other: as Mary Hawkesworth notes, the implicit "rules" of discourse for members of parliament in Britain, Canada, and Australia can be characterized as "loud, aggressive, and combative" and can include "screaming, shouting, and sneering that can create no-win situations for women members. Women who adopt this combative style are ridiculed and patronized by their male counterparts, whereas women who

[37] I use the word "social power" here in Keith Dowding's sense, as that is the kind of power Young seems to be referring to, when she speaks of "power over others by means of mobilizing threats of sanction or offers of desired goods"; see I.M. Young, *Structure as the Subject of Justice*, op. cit., p. 61. Dowding's concept of "social power" includes the ability not just to threaten but to persuade A, such that A changes their preference structure to bring about an end that is different to that of A's initial preference structure. See K. Dowding, *Encyclopedia of Power*, SAGE, Thousand Oaks 2011, pp. 616–619.

[38] In the opening paragraph of *Justice and the Politics of Difference*, she declares social groups as the focus of her philosophical inquiry and then in *Equality of Whom? Social Groups and Judgments of Injustice*, she challenges the assumption "that the units we should be comparing when we make judgments of inequality are individuals"; see I.M. Young, *Equality of Whom? Social Groups and Judgments of Injustice*, "The Journal of Political Philosophy" 2001, Vol. 9, No. 1, pp. 1–18; and I.M. Young, *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990, p. 3.

[39] I.M. Young, *Justice and the Politics of Difference*, op. cit., p. 14.

opt for a more demure, consultative, and collaborative style are labelled 'weak' or 'unfit' for the job."[40]

In her earlier work, *Justice and the Politics of Difference*, Young argues that sexism and other -isms occur at the level of practical consciousness – in the aversive (perhaps unintended) reactions one might have to the Other, including sexist acts,[41] homophobia,[42] ageism and ableism,[43] and racism.[44] Insofar as Giddens's notion of practical consciousness is tied to unverbalizable rule-following, I take Young to mean that these aversive sexist (and so on) reactions are themselves the silent enactment of certain "group-focused routines."[45] This is what can be understood when she says that racism etc. is "enacted in [US] society […] in informal, often unnoticed and unreflective speech, bodily reactions to others, conventional practices of everyday interaction and evaluation, aesthetic judgments, and the jokes, images, and stereotypes pervading the mass media."[46]

By the time she wrote *Structure as the Subject of Justice*, Young seems to have "add[ed] some dimensions"[47] to this – in particular, Bourdieu's notion of "habitus," wherein bodily comportments, reactions, tastes, and preferences – stratified by class, wealth, and other socially salient categorizations – silently signal one's social position to others in such forms as voice, gesture, and a preference for, for example, scotch over beer (or vice versa). This represents an important complement to her account of -isms, showing how habitus, for example in the form of one's desire to find an apartment in a (white) middle-class neighbourhood ("where others like me live"), "(unconsciously) operates to reproduce structural inequalities" – where "structural inequalities" refer to "categorical inequalities, typically along the lines of class or class fraction, race, gender, ability, and sometimes ethnicity."[48]

The potential for social liberation from -isms – that is, for the elimination within contemporary society of racism, sexism, and so on – is available via Young's adoption of Giddens's conceptual tool of structuration. For much of

40 K. Dowding, *Encyclopedia of Power*, op. cit., p. 255.
41 I.M. Young, *Justice and the Politics of Difference*, op. cit., p. 133.
42 Ibid., p. 146.
43 Ibid., p. 147.
44 Ibid., p. 151.
45 Ibid., p. 146.
46 Ibid., p. 148.
47 See I.M. Young, *Structure as the Subject of Justice*, op. cit., p. 62.
48 Ibid., p. 59.

the 20th century, social theorists had tended to coalesce around either an agent-centric paradigm, wherein individual actions are conceptualized as autonomous and largely unconstrained, or one that is structure-centric, in which structures of power constrain/determine human behaviour. Giddens's structuration, on the other hand, recognizes a duality: structure shapes human action, yet it is simultaneously and recursively constructed by those human actions. It is this latter aspect that suggests that humans have the capacity to alter their actions, to change their behaviours – including those actions and behaviours that reproduce -isms at the level of practical consciousness. Of course, the fact that they play out inadvertently poses a challenge: if humans are unaware of their racist, sexist, etc., tendencies, how can they correct them? The solution is to raise consciousness, to bring that which is inadvertent to the level of discursive consciousness – where discursive consciousness is understood as a level of experience where actors know what they are doing and can provide reasons for their behaviour. Consciousness-raising happens through social groups gathering to discuss their experiences of being treated as the Other – with recognition of common themes shared across their experience, and a vocabulary with which to describe it, emerging in their discussions. That occurred with the women's movement in the 1960s, and with the Black liberation movement in the late 1960s. Miranda Fricker provides an excellent example that sheds light on how consciousness-raising works: when women endured sexualized comments in the workplace etc., prior to the 1960s it was brushed aside as "flirting" or "harmless fun"; but when several women came together to discuss similar experiences, they began to develop a vocabulary around it – calling it "sexual harassment" – and eventually bringing/raising awareness of the wrongness of such behaviour to the level of men's discursive consciousness.[49]

Let us take stock. It should be apparent, at this point, that Young's critical theory grounds the justness or injustice of a social arrangement/outcome in a social theory of structural power. She provides a suite of concepts and tools that philosophers could draw upon to normatively reason about socially unjust outcomes – including, structural power; social groups and their experience of -isms; practical consciousness; consciousness-raising activities; and Giddens's structuration. Crucially, the incorporation of consciousness-raising spaces within her framework provides the mechanism for racist, sexist, etc., behaviours to be "named" – for example,

[49] M. Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, Oxford 2007, pp. 150–151, https://doi.org/10.1093/acprof:oso/9780198237907.001.0001.

as "sexual harassment," as we saw in Fricker's account. Left unnamed, they go undetected – and the behaviours continue unabated, re-enacted in society (as was the case with the inappropriate, sexualized comments that were part of workplace culture before women's groups called them out, and male managers were sent to workplace gender-awareness workshops). And, insofar as such behaviours are re-enacted in society, they are more likely to be reproduced in AI outcomes. It is worth highlighting the robustness of Young's account. Not only does it give us the concepts with which to ideate racism, sexism, and other -isms – it also provides the tool that helps counter (and perhaps one day eliminate) them.

Furthermore, her consciousness-raising spaces can nourish the moral deliberations of philosophers. When morally relevant facts rise to discursive consciousness, philosophers have a broader array of facts to contemplate. Additionally, they gain the capacity to censure the behaviour of any perpetrators who, though now aware, are nonetheless unmoved.

## 5. AI Outcomes

I have shown that, when a social outcome implicates racism, sexism, and other -isms, an assessment of its injustices necessitates the use of a critical theory and an account of structural power. However, insofar as this approach tackles -isms *tout court*, advocacy for it would seem to hold even without AI.

Does anything change when we apply the approach to AI? Certainly, AI compounds the issue, reproducing those -isms in its outputs. Further, given the opacity of neural networks, we might not understand why that has happened (at least at the level of/inside the black box). However, the social theory within Young's account allows us to better understand the social phenomena that caused the racist, sexist, etc., AI outputs. As we have seen, an integral part of Young's critical theory is the value it gives to consciousness-raising spaces. Insofar as consciousness-raising helps curtail inadvertent sexism, racism, etc., and insofar as those -isms are moral wrongs that ought to be curtailed, it follows that consciousness-raising spaces ought to be developed and maintained to help identify and address instances of AI bias.

But how would consciousness-raising activities help here? How would they ameliorate the detection and redress of AI bias? Such bias sometimes only comes

to light when historically marginalized people have a "hunch" that the algorithm is treating them differently. Without consciousness-raising activities, that hunch may remain undisclosed; it may even remain unidentified as a phenomenon – silently and unwittingly endured by marginalized people as "an inconvenience," rather than a form of "discrimination" or "harassment."[50] Consciousness-raising activities, on the other hand, provide a forum for discussing such hunches, sharing adverse experiences, and identifying patterns of AI bias. This process allows for the feedback of identified bias to AI developers. For example, African American and Hispanic communities could discuss the impacts of predictive policing and parole review AI systems on their lives; by sharing their individual experiences of (what at first may seem like) "unfortunate" parole denials, a pattern becomes discernible and (racial) bias becomes apparent.

One important question to consider is whether consciousness-raising activities replace existing mechanisms for addressing AI fairness, or do they complement them. Consider some existing mechanisms for addressing AI fairness:

- COMPAS, an AI tool used to predict recidivism, was shown to be biased against Black offenders – prompting the development of a race-neutral version of the algorithm.[51]
- Some companies deploy "gender decoders" to analyse job descriptions and detect subtle language biases that may deter women from applying – terms like "executes" or "competitive" might be flagged as masculine-coded.[52]
- To counteract the over-representation of certain groups in training data, re-sampling techniques may be used to ensure more balanced representation – as seen with facial recognition technologies.

Whilst these existing mechanisms may be effective to some extent, consciousness-raising activities can enhance their effectiveness by alerting AI developers to instances of AI bias and the need for such interventions.

---

[50] This is analogous to the experience for many women in the 1950s who faced inappropriate, sexualized behaviour from male colleagues. At the time, such behaviour was often dismissed as "flirting" and considered an "inconvenience" by some female colleagues. It was only later, through consciousness-raising activities and the sharing of experiences, that they came to recognize and identify these behaviours as "discrimination" and "sexual harassment."

[51] J. Angwin et al., *Machine Bias*, in: *Ethics of Data and Analytics: Concepts and Cases*, ed. K. Martin, Auerbach Publications, Boca Raton 2016, pp. 254–264.

[52] K. Crawford, T. Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets*, "AI & Society" 2021, Vol. 36, pp. 1105–1116, https://doi.org/10.1007/s00146-021-01162-8.

That said, some existing AI fairness mechanisms face legal constraints because they often require access to sensitive attributes (such as gender or ethnicity) that privacy laws may ringfence.[53] In this context, consciousness-raising activities could offer a viable alternative. Instead of mining sensitive data to detect bias or demonstrate compliance with fairness standards, AI developers can engage with discursive consciousness-raising forums. These forums bring attention to biases related to gender, ethnicity, and other protected traits, allowing developers to identify issues through participant feedback[54] rather than through direct access to sensitive information.

Let's consider a potential objection to the analysis presented in this paper. The paper explored two possible approaches to appraising the justness or injustice of AI outcomes: Rawls's and Young's. A critic might ask: there are other abstract theories within political philosophy other than Rawls's – why consider his? Our answer is twofold. First, that we can't not consider him. His theory has become the dominant ideal theory in political philosophy over the past 50 years, shaping the thinking of many contemporary political philosophers. By engaging with Rawls, we interact with how a substantial portion in the field approach questions of justice and injustice. Second, AI scholars have already reached for Rawls's theory to answer questions about AI-exacerbated social injustice. Indeed, as Jørgensen and Søgaard note, "Researchers and industry developers in artificial intelligence (AI) and natural language processing (NLP) have uniformly adopted a Rawlsian definition of fairness."[55] One reason I assessed Rawls's theory within this paper was to make clear that it cannot answer the sorts of questions that worry many who study AI bias. Our analysis is intended to save them time and

[53] Yan et al. make this point too; see S. Yan, H.-T. Kao, E. Ferrara, *Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes*, "Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency" 2020, p. 1715, https://doi.org/10.1145/3340531.3411980.

[54] Of course, a form of feedback collection already exists within AI: a user may be presented with a short, in-app survey or with a prompt to rate the fairness of the app; an app may include a "Report bias" button; or the AI system might monitor user behaviour, noting that the user frequently overrides AI recommendations. But, whereas these existing mechanisms entail feedback from a single user, the consciousness-raising feedback is from many users, who, through the process of communicating their shared experience with one another, have clarified the phenomenon of group bias.

[55] A.K. Jørgensen, A. Søgaard, *Rawlsian AI Fairness Loopholes*, op. cit., p. 1185.

effort – steering them away from a philosophical path that cannot speak to the issues they seek to tackle within AI fairness.

Consider a second query about the paper. The critic might acknowledge that Young's theory indeed considers structural power, but then ask: but why should we? At one level, we can respond that, unless we do, we cannot grapple with those AI outcomes that implicate and reproduce structural power inequalities. But let's consider the critic's query more deeply. Perhaps they are saying that, insofar as structural racism, sexism, etc., are inadvertent, we cannot assign moral blame/culpability to anyone for them – as such, we should ignore -isms in our deliberations about the moral permissibility of AI outcomes. My response is that this suggestion fails to grasp the interplay between Giddens's notion of structuration and the revelatory effects of consciousness-raising activities. The latter provides actors with information and insights that allow them to recognize their actions and reflect on them. The former shows us that agents retain agency – they *can* change their actions; and insofar as persons can change a morally impermissible or unjust action, we can hold them responsible – indeed, we could blame them, even (once we conduct appropriate moral deliberations that weigh any mitigating factors that could account for their inaction).[56]

## 6. Conclusion

Many scholars have engaged with Rawls's justice as fairness when studying AI fairness. We showed, though, that Rawls's theory, lacking a sociological theory of structural power, was not fit for that purpose – but that it was never intended for that purpose, either: it is supposed to move us towards Rawls's ideal version of justice, rather than to address, and move us away from, any particular -ism

[56] Tetyana Krupiyu argues that we ought to recognize the computer/data scientist's contribution to AI, rather than just thinking of the algorithm and its outputs, since this helps "capture the fact that computer scientists make subjective decisions in the course of creating the architecture that enables the AI decision-making process to collect, aggregate and analyse data. […] Often, the decisions of computer scientists are hidden and reflect a particular understanding of the world. For example, computer scientists make assumptions when deciding how to represent a person in a model" (T. Krupiyu, *A Vulnerability Analysis: Theorising the Impact of Artificial Intelligence Decision-Making Processes on Individuals, Society and Human Diversity from a Social Justice Perspective*, "Computer Law & Security Review" 2020, Vol. 38, 105429, https://doi.org/10.1016/j.clsr.2020.105429, p. 8 of 25).

injustices. This revelation allowed us to conclude that AI ethicists should not look to Rawls when they ask questions about AI decisions that are racist, sexist, etc.

On the other hand, we showed that Young's approach, drawing on a sociological theory of structural power, is well-suited to the task. Her concept of practical consciousness, as we saw, accounted for unspoken, pernicious aspects of racism, sexism, and so on. Moreover, Young's device of consciousness-raising activities, as I showed, can illuminate and "name" unjust behaviours. That can, as I argued, nourish philosophers' moral reasoning about AI outcomes that are racist, sexist, etc. It can, also, as we saw, help remove the racism, sexism, and other -isms that get reproduced in AI outcomes.

## Bibliography

Anderson E., *What Is the Point of Equality?*, "Ethics" 1999, Vol. 109, No. 2, pp. 287–337.

Angwin J., Larson J., Mattu S., Kirchner L., *Machine Bias*, in: *Ethics of Data and Analytics: Concepts and Cases*, ed. K. Martin, Auerbach Publications, Boca Raton 2016, pp. 254–264.

Barsotti F., Koçer R.G., *MinMax Fairness: From Rawlsian Theory of Justice to Solution for Algorithmic Bias*, "AI & Society" 2024, Vol. 39, pp. 961–974, https://doi.org/10.1007/s00146-022-01577-x.

Bay M., *Participation, Prediction, and Publicity: Avoiding the Pitfalls of Applying Rawlsian Ethics to AI*, "AI and Ethics" 2024, Vol. 4, pp. 1545–1554, https://doi.org/10.1007/s43681-023-00341-1.

Binns R., *Algorithmic Accountability and Public Reason*, "Philosophy and Technology" 2018, Vol. 31, pp. 543–556.

Buolamwini J., Gebru T., *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, "Proceedings of Machine Learning Research" 2018, Vol. 81, pp. 1–15.

Buranyi S., *Rise of the Racist Robots: How AI Is Learning All Our Worst Impulses*, "The Guardian," 8.08.2017, URL: https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

CBC Radio, *Police Are Considering the Ethics of AI, Too*, 21.09.2018, URL: https://www.cbc.ca/radio/spark/tech-in-policing-1.4833189/police-are-considering-the-ethics-of-ai-too-1.4833194.

Crawford K., Paglen T., *Excavating AI: The Politics of Images in Machine Learning Training Sets*, "AI & Society" 2021, Vol. 36, pp. 1105–1116, https://doi.org/10.1007/s00146-021-01162-8.

Daniels N., ed., *Reading Rawls: Critical Studies on Rawls' A Theory of Justice*, Stanford University Press, Stanford 1975.

Dastin J., *Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women*, Reuters, 9.10.2018, URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Dowding K., *Encyclopedia of Power*, SAGE, Thousand Oaks 2011.

Dunn E., *Public Attitudes to Data and AI: Tracker Survey*, Centre for Data Ethics and Innovation, London 2022.

Farrelly C., *Justice in Ideal Theory: A Refutation*, "Political Studies" 2007, Vol. 55, pp. 844–864.

Ferrara E., *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies*, "Sci" 2024, Vol. 6, No. 1, pp. 1–15, https://doi.org/10.3390/sci6010003.

Fricker M., *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, Oxford 2007, https://doi.org/10.1093/acprof:oso/9780198237907.001.0001.

Gabriel I., *Toward a Theory of Justice for Artificial Intelligence*, "Daedalus" 2022, Vol. 151, No. 2, pp. 218–231, https://doi.org/10.1162/daed_a_01911.

Giddens A., *Central Problems in Social Theory: Action, Structure, and Contradiction in Social Analysis*, MacMillan Education, London 1979.

Giddens A., *The Constitution of Society: Outline of the Theory of Structuration*, Polity, Cambridge 1986.

Heidari H., Ferrari C., Gummadi K., Krause A., *Fairness behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making*, "Advances in Neural Information Processing Systems" 2018, Vol. 31.

Jørgensen A.K., Søgaard A., *Rawlsian AI Fairness Loopholes*, "AI and Ethics" 2022, Vol. 3, pp. 1185–1192, https://doi.org/10.1007/s43681-022-00226-9.

Krupiy T., *A Vulnerability Analysis: Theorising the Impact of Artificial Intelligence Decision-Making Processes on Individuals, Society and Human Diversity from a Social Justice Perspective*, "Computer Law & Security Review" 2020, Vol. 38, 105429, https://doi.org/10.1016/j.clsr.2020.105429.

MacMillan D., *Eyes on the Poor: Cameras, Facial Recognition Watch over Public Housing*, "The Washington Post," 16.05.2023, URL: https://www.washingtonpost.com/business/2023/05/16/surveillance-cameras-public-housing/.

McNay L., *Recognition as Fact and Norm: The Method of Critique*, in: *Political Theory: Methods and Approaches*, eds. D. Leopold, M. Stears, Oxford University Press, Oxford 2008, pp. 85–105.

Mills C.W., *"Ideal Theory" as Ideology*, "Hypatia" 2005, Vol. 20, No. 3, pp. 165–184.

Mills C.W., *Retrieving Rawls for Racial Justice? A Critique of Tommie Shelby*, "Critical Philosophy of Race" 2013, Vol. 1, No. 1, pp. 1–27.

PBS Frontline, *A Class Divided*, "CosmoLearning" 1985.

Rafanelli L.M., *Justice, Injustice, and Artificial Intelligence: Lessons from Political Theory and Philosophy*, "Big Data and Society" 2022, Vol. 9, No. 1, https://doi.org/10.1177/20539517221080676.

Rawls J., *The Law of Peoples*, Harvard University Press, Cambridge, MA, 1999.

Rawls J., *A Theory of Justice*, Harvard University Press, Cambridge, MA, 1971.

Reuters, *Amazon Ditched AI Recruiting Tool that Favored Men for Technical Jobs*, "The Guardian," 11.10.2018, URL: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine.

Simmons A.J., *Ideal and Nonideal Theory*, "Philosophy & Public Affairs" 2010, Vol. 38, pp. 5–36.

Weidinger L., McKee K., Everett R., Huang S., Zhu T., Chadwick M., Summerfield C., Gabriel I., *Using the Veil of Ignorance to Align AI Systems with Principles of Justice*, "Proceedings of the National Academy of Sciences of the United States of America" 2023, Vol. 120, e2213709120, https://doi.org/10.1073/pnas.2213709120.

Yan S., Kao H.-T., Ferrara E., *Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes*, "Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency" 2020, pp. 1715–1724, https://doi.org/10.1145/3340531.3411980.

Young I.M., *Communication and the Other: Beyond Deliberative Democracy*, in: *Democracy and Difference: Contesting the Boundaries of the Political*, ed.

S. Benhabib, Princeton University Press, Princeton 1996, pp. 120–136, https://doi.org/10.1515/9780691234168-007.

Young I.M., *Equality of Whom? Social Groups and Judgments of Injustice*, "The Journal of Political Philosophy" 2001, Vol. 9, No. 1, pp. 1–18.

Young I.M., *Justice and the Politics of Difference*, Princeton University Press, Princeton 1990.

Young I.M., *Structure as the Subject of Justice*, in: I.M. Young, *Responsibility for Justice*, Oxford University Press, Oxford 2011, https://doi.org/10.1093/acprof:oso/9780195392388.003.0002.

Zhang M., *Google Photos Tags Two African-Americans as Gorillas through Facial Recognition Software*, "Forbes," 1.07.2015, URL: https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/.

Edukacja Filozoficzna
ISSN 0860-3839, eISSN 2956-8269
DOI: 10.14394/edufil.2025.0005
ORCID: 0009-0003-5278-6965

# A Philosophical Account of Shared Autonomy and Moral Agency in Human–AI Teams

Max Parks
(University of Michigan, Mott Community College)

**Abstract:** This paper develops a framework for understanding autonomy and moral agency in hybrid human–AI systems. We begin with an examination of the autonomous vehicle "trolley problem," a problem for how the AI in autonomous vehicles should be programmed. This scenario reveals a fundamental distinction between computational reasoning, where AI excels, and social-moral judgement, where human capabilities remain essential. The autonomous vehicle scenario exemplifies broader challenges in human–AI collaboration. Purely computational approaches to moral decisions prove insufficient, as they lack the social understanding and attentive care characteristic of human judgement. This insufficiency becomes particularly apparent in applications attempting to replicate human social relationships, where the absence of what Ellen Ullman in her article *Programming the Post-Human: Computer Science Redefines "Life"* on posthumanism terms genuine "presence" and mutual recognition creates risks of diminishing rather than enhancing human capabilities. By examining these cases, this paper develops principles for responsible integration of AI capabilities while preserving meaningful human agency.

**Key words:** agency, AI, shared, autonomy, team, distributed, emergent, moral

## 1. Introduction*

As artificial intelligence (AI) systems increasingly perceive, decide, and act alongside us, agency is no longer the property of a single rational subject. Consider the cases of autonomous vehicles that decide whether to swerve into pedestrians; social robots that promise unconditional companionship; and chatbots that counsel teenagers in distress. In such cases, action is distributed across biological beings and computational artefacts whose capacities are neither identical nor interchangeable. Most analyses respond by asking which component "really"

* Max Parks would like to give credit to and thank Mark Allison (University of Michigan, Flint) for the images included in this paper.

makes the choice or which optimization rule should be encoded. While AI systems can calculate probable outcomes with precision, they lack what Ellen Ullman identifies as authentic presence: the capacity for genuine moral understanding and social recognition that characterizes human moral judgement.[1] Moral life originates not in detached calculation but in relations of care, the networks of attention, dependency, and mutual recognition through which human beings sustain one another.[2]

Standard approaches in AI ethics find the correct decision rule, embed it in software, and verify compliance. That works adequately for narrowly technical harms (for example, data leakage), but it fails in situations where the quality of attention and responsiveness is itself the morally salient variable. A self-driving car that minimizes expected fatalities may still wrong its passenger if the passenger never consented to being sacrificed, just as a companion robot that recognizes and responds to a lonely elder's mood may still erode her well-being by displacing human contact. Neither outcome registers as a violation within purely utilitarian or deontological spreadsheets, yet both reflect a failure to honour the vulnerability and relational needs of the people involved.

Feminist ethics of care offers a vocabulary built precisely for these failures. Care theorists begin from the fact of universal dependence: all persons spend portions of their lives relying on the skill and goodwill of others. Moral agency therefore consists in *attending to, interpreting, and meeting concrete needs* within asymmetric relationships.[3] Care is neither sentimental attachment nor unpaid domestic labour; it is a socio-material practice marked by attentiveness, responsibility, competence, and responsiveness.[4] From this standpoint, the central issue about AI and agency is not whether machines can become moral agents but whether their deployment enlarges or diminishes the practices through which people recognize and satisfy one another's needs.

---

1 E. Ullman, *Programming the Post-Human: Computer Science Redefines "Life"*, "Harper's Magazine" 2002, Vol. 305(1829), pp. 60–70.

2 V. Held, *The Ethics of Care: Personal, Political, and Global*, Oxford University Press, Oxford 2006; J. Tronto, *Caring Democracy: Markets, Equality, and Justice*, New York University Press, New York 2013.

3 N. Noddings, *Caring: A Relational Approach to Ethics and Moral Education*, 2nd ed., University of California Press, Berkeley 2013; E.F. Kittay, *Love's Labor: Essays on Women, Equality, and Dependency*, Routledge, New York 1999.

4 J. Tronto, *Caring Democracy*, op. cit.

A complementary strand, relational autonomy, sharpens the point. Autonomy is not the self-sufficient exercise of will but an achievement realized through social recognition and answerability.[5] If an AI-mediated decision leaves no recognizable human capable of apologizing, explaining, or repairing harm, relational autonomy, and thus moral legitimacy, is compromised even if aggregate utility rises.

This paper advances a single aim: to develop a care-centric conceptual and normative framework for hybrid human–AI agency, and to demonstrate its practical value through two flagship cases, autonomous vehicles and social robots. Rather than treating care as an add-on to existing control paradigms, we place it at the centre of analysis, focusing on who is recognized and attended to, how capacity for relational self-direction is preserved or eroded, and how accountability lines are maintained.

We focus on autonomous vehicles and social robots because together they span the continuum from high-stakes physical risk to relational and affective risk, and both have robust public datasets that allow fine-grained care analysis. Section 2 situates care ethics and relational autonomy against traditional control-centric theories and explains how technology should instead be evaluated by how it contributes to or facilitates caring relationships. Section 3 applies the framework to autonomous-vehicle crash scenarios and to therapeutic versus companion social robots, showing how caring relations are sustained or undermined in each domain. Section 4 covers a Care-Impact Assessment template. Section 5 addresses the many-hands problem, mapping legal responsibility and regulatory instruments onto care chains in both cases. Section 6 concludes by outlining a research agenda for AI development that keeps caring presence and relational accountability at its core.

By foregrounding care rather than control, we argue, designers and policymakers can spot ethical failures invisible to optimization metrics, address hidden inequities in labour and risk, and build hybrid systems that genuinely enhance rather than erode human well-being.

[5] C. Mackenzie, N. Stoljar, eds., *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, Oxford University Press, New York 2000.

## 2. Agency in Hybrid Human–AI Teams

### 2.1. Why Traditional Agency Accounts Falter in Hybrid Settings

Most discussions of machine autonomy inherit an implicit picture from classic action theory: a single rational subject forms an intention, issues motor commands (or code), and bears responsibility for the outcome.[6] When AI enters the loop, scholars typically tweak only the locus of control, asking whether the human still "pulls the lever" or whether the algorithm does. This control-centric focus abstracts away the relational context of action. A collision-avoidance algorithm may prevent bodily harm, yet neglect to honour a passenger's legitimate expectation of having her safety prioritized. Control theory registers only event-level success or failure, not the relational meaning of those outcomes. Attempts to patch control-centric ethics by adding preference retrieval, meta-utility functions, or "ethical governors" fail to resolve these omissions because the omissions are structural, not parametric. We need an alternative starting point.

### 2.2. Feminist Ethics of Care and Relational Autonomy

*Caring presence.* For Virginia Held, the founding act of care is *attentiveness*: noticing another's need in its concrete particularity.[7] The moral failure in many AI misfires is not malice or mis-optimization but *inattention*, with no one present who can see and respond.

*Dependency networks.* Eva Feder Kittay emphasizes that every individual, no matter how empowered, participates in chains of dependency.[8] Children, the ill, and the elderly rely more heavily on caregivers, and caregivers, in turn, depend on wages, social recognition, and respite. When AI systems replace some nodes in these chains, the *structure* of dependency shifts, often invisibly. Relatedly, care theory is also concerned with whether deployment of an AI system reinforces, redistributes, or remediates existing axes of domination, suggesting that we map who gains free time, whose labour is displaced, and whose safety is prioritized.[9] For example, autonomous-vehicle risk externalities often fall

[6] A.R. Mele, *Motivation and Agency*, Oxford University Press, Oxford 2003.
[7] V. Held, *The Ethics of Care*, op. cit.
[8] E.F. Kittay, *Love's Labor*, op. cit.
[9] N. Bahrami, *AIgemony: Power Dynamics, Dominant Narratives, and Colonisation*, "AI and Ethics" 2025, Vol. 5, pp. 5081–5103, https://doi.org/10.1007/s43681-025-00734-4.

on non-driver road users, such as pedestrians, cyclists, gig-economy couriers, groups already under-served by city infrastructure.

*Relational accountability*. Catriona Mackenzie and Natalie Stoljar argue for an account of autonomy as the capacity to live according to values and projects recognized and supported by others.[10] Accountability, in this view, is not just causal responsibility but *answerability*, the ability to justify one's actions to those affected. An opaque optimization routine that sacrifices a passenger severs this line of answerability.

Whenever we later ask whether an autonomous vehicle or social robot behaves ethically, we check (a) whether someone or something is attentively present to concrete need; (b) how the system reshapes dependency networks; and (c) whether those affected can hold a recognizable agent to account. The empirical and regulatory analyses in sections 3–5 all map directly onto this triad.

Having set out the three background assumptions – caring presence, dependency networks, and relational accountability – we still need a way to trace how those values are applied in practice. Joan Tronto's procedural account of care does precisely this, breaking the practice into four successive phases.[11]

1. Caring *about* (attentiveness) – sensors detect hazard but may not register social meaning (for example, stroller versus shopping cart).
2. Caring *for* (responsibility) – who is tasked to intervene: the passenger, remote operator, or original equipment manufacturer?
3. Care *giving* (competence) – does the AI system possess the skills to meet the need without degrading human skills?
4. Care *receiving* (responsiveness) – can those affected signal satisfaction or distress back into the loop?

Taken together, the four phases give us a step-by-step checklist for evaluating care in practice: first ask who notices need, then who takes responsibility, whether the system is competent to meet that need, and finally whether those affected can signal satisfaction or distress back into the loop. For example, full self-driving AI disengagements fail phase 2 (responsibility) when drivers over-trust automation, and companion robots often fail phase 4 when users cannot register loneliness once the novelty fades.

[10] C. Mackenzie, N. Stoljar, eds., *Relational Autonomy*, op. cit.

[11] J. Tronto, *Caring Democracy*, op. cit.

## 2.3. "Care Prosthesis" Metaphor

Andy Clark and David Chalmers famously argue that notebooks or smartphones can become non-biological parts of cognition when they integrate seamlessly into task routines.[12] Adopting this insight, we propose that AI modules function ethically when they act as care prostheses, or tools that enhance the caregiver's capacity for attentiveness, responsibility, competence, and responsiveness, without eclipsing the relational practice itself.

For example, an autonomous-vehicle perception stack that detects a cyclist in a driver's blind spot extends attentiveness. But if the same system unilaterally executes a passenger-sacrifice trajectory without soliciting consent, it strips the human of relational accountability. The same hardware can either augment or erode care, depending on how it is programmed and used.

The prosthesis metaphor imposes a normative limit: a prosthetic limb is valuable because it restores agency to the person, not because it can walk away on its own. Likewise, AI should restore or enhance human caring relations, but when it claims authority to replace those relations entirely, it crosses the ethical line.

Figure 1 brings the theoretical strands together. Only where computational capability is integrated with human attentiveness and a channel for relational accountability do we obtain genuine shared autonomy.

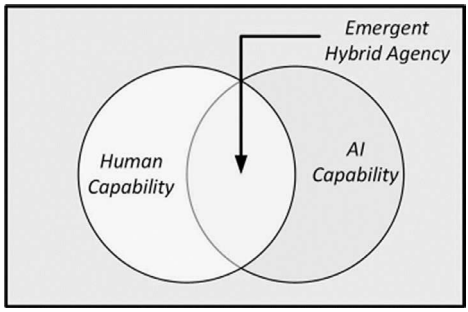



Figure 1: Emergent agency in human-AI teams
Source: Mark Allison.

[12] A. Clark, D.J. Chalmers, *The Extended Mind*, "Analysis" 1998, Vol. 58, No. 1, pp. 7–19, https://doi.org/10.1093/analys/58.1.7.

## 3. Autonomous Vehicles: Crash Scenarios and the Politics of Caring Presence

The autonomous vehicle confronting the trolley problem, choosing between protecting its passenger or multiple pedestrians,[13] serves as a paradigmatic case for examining the limitations of purely computational approaches to moral decisions. Long before the advent of self-driving cars, the trolley problem originated in philosophical discussions of moral principles and obligations.[14] Initially, the problem asked whether it is permissible to pull a lever, redirecting a trolley onto a track that would kill one person to save five others. Philosophers use these scenarios to test moral intuitions about permissible harm, double effect, and the difference between killing versus letting die.

With the rise of autonomous vehicle technologies, the trolley problem became a practical design concern, as engineers and ethicists alike wonder how to program vehicles to respond in collision scenarios where fatalities may be unavoidable. Maximilian Geisslinger et al. reject pure utilitarian or deontological approaches, instead advocating for an "ethics of risk" framework that combines three principles: minimizing overall risk, ensuring equality in risk distribution, and protecting the worst-off.[15] They argue this provides a better way to handle inevitable uncertainty in driving scenarios. Chiara Lucifora et al.'s experimental study reveals an important gap between "hot" immediate moral decisions made while driving versus "cold" deliberative choices made with time to reflect.[16] Their findings suggest that while people tend towards utilitarian choices in the moment, they incorporate broader moral considerations like family values and social roles when given time to deliberate; however, it is not obvious how this should inform autonomous-vehicle programming.

---

[13] S. Nyholm, J. Smids, *The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?*, "Ethical Theory and Moral Practice" 2016, Vol. 19, pp. 1275–1289, https://doi.org/10.1007/s10677-016-9745-2.

[14] P. Foot, *The Problem of Abortion and the Doctrine of Double Effect*, "Oxford Review" 1967, Vol. 5, pp. 5–15; J.J. Thomson, *Killing, Letting Die, and the Trolley Problem*, "The Monist" 1976, pp. 204–217.

[15] M. Geisslinger et al., *Autonomous Driving Ethics: From Trolley Problem to Ethics of Risk*, "Philosophy & Technology" 2021, Vol. 34, No. 4, pp. 1033–1055.

[16] C. Lucifora et al., *Moral Dilemmas in Self-Driving Cars*, "Rivista Internazionale di Filosofia e Psicologia" 2020, Vol. 11, No. 2, pp. 238–250, https://doi.org/10.4453/rifp.2020.0015.

### 3.1.1. Technical Context and Empirical Record

In March 2018 an experimental Uber test vehicle operating in "computer control" mode struck and killed a pedestrian in Tempe, Arizona. The US National Transportation Safety Board (NTSB) found that the perception stack identified her six seconds before impact yet re-classified her several times and, by design, suppressed emergency braking unless the safety driver intervened. The driver was not paying adequate attention.[17]

The baseline autonomous-vehicle pipeline from perception to trajectory planning operates on millisecond cycles. It excels at kinematic optimization but knows nothing of social or moral meaning; a child and a rolling trash can may both appear as "dynamic obstacles." Manufacturers sometimes propose "ethical algorithms" that minimize statistically expected fatalities, but we will explore in detail why caring is a necessary condition to include in the decision-making process.[18]

### 3.1.2. Care Analysis

Sensors detected the pedestrian, but no agent in the loop noticed a vulnerable person in need of care. The system's cost-function logic suppressed braking to avoid false positives, and the safety driver's visual attention was divided. The failure illustrates Held's claim that moral breakdown often begins with inattention rather than ill-will.[19]

NTSB concluded that Uber Advanced Technologies Group's "inadequate safety culture" contributed to the pedestrian's death. But with responsibility dispersed across software engineers, safety operators, and state regulators, we have an instance of the many-hands problem.[20] Care theory would ask: "Which party was positioned to recognize the pedestrian's need and respond competently?" The answer, in this case, was *no one*. Machine perception can out-perform humans at night-time object detection, yet it lacks the moral competence of interpreting

[17] National Transportation Safety Board, *Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian*, URL: https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx.

[18] J.-F. Bonnefon, A. Shariff, I. Rahwan, *The Trolley, the Bull Bar, and Why Engineers Might Fear Ghosts: An Empirical Study of Morally Loaded Technical Decisions*, "Proceedings of the IEEE" 2019, Vol. 107, No. 3, pp. 502–504, https://doi.org/10.1109/JPROC.2019.2897447.

[19] V. Held, *The Ethics of Care*, op. cit.

[20] I. van de Poel, *The Problem of Many Hands*, in: I. van de Poel, L. Royakkers, S.D. Zwart, *Moral Responsibility and the Problem of Many Hands*, Routledge, New York 2015, pp. 50–92.

a cyclist walking a bike as a special vulnerability category. Neither the algorithm nor any Uber executive could apologize in person. Relational autonomy deems such absence of *answerability* a secondary harm.[21]

### 3.1.3. The Utilitarian Temptation and Its Care-Ethics Limits

Proponents of utilitarianism argue that autonomous vehicles should simply minimize overall harm, even if passengers must be sacrificed.[22] Large-scale Moral Machine surveys show abstract public support for such rules.[23] Yet researchers such as Lucifora and colleagues found that under time pressure, drivers in simulator experiments revert to passenger-protective instincts.[24] From a care standpoint, the utilitarian proposal fails on two counts:

1. Relational accountability. A passenger never asked to die for statistical strangers; sacrificing her without prior assent severs answerability lines. Nel Noddings would label this a failure to maintain caring presence for the passenger.[25]
2. Asymmetric burdening. Passengers disproportionately bear risk, while system designers avoid bodily harm themselves, a distribution incompatible with Tronto's democratic care ideal.[26]

Given these considerations, it seems a care-centric redesign facilitating care-based decisions requires the system to complement a user's capacity to care, so for example, notifying the passenger early and requesting a policy preference (for example, "protect occupants," "minimize harm overall," or "driver decides in real time"). This would serve to complement or enhance caring human presence.

Focusing on care also means adopting transparent UX practices, such as having risk trade-offs displayed in everyday language ("In this route, a severe crash is one in 10 million; here is how pedestrians' risk compares to yours"). This would maximize the contributions of both parties, that is, the information provided by the AI system and the human counterpart using that information to make informed judgement calls.

---

21 C. Mackenzie, N. Stoljar, eds., *Relational Autonomy*, op. cit.

22 J.F. Bonnefon, A. Shariff, I. Rahwan, *The Trolley, the Bull Bar, and Why Engineers Might Fear Ghosts*, op. cit.

23 E. Awad et al., *The Moral Machine Experiment*, "Nature" 2018, Vol. 563, pp. 59–64.

24 C. Lucifora et al., *Moral Dilemmas in Self-Driving Cars*, op. cit.

25 N. Noddings, *Caring: A Relational Approach to Ethics and Moral Education*, op. cit.

26 J. Tronto, *Caring Democracy*, op. cit.

Lastly, to respect relational accountability, a care-centred design should allow event data to be logged so a human stakeholder can explain and, if needed, initiate changes and apologize.

The autonomous-vehicle case shows how caring presence can vanish when relational responsibility is neglected in favour of optimizing algorithms to operate without the caring presence of a human agent. Only by embedding such structures can an autonomous-vehicle system extend, rather than erode, the relational fabric of road safety. To be clear, not every real-world episode fits the failure narrative, as automation can unobtrusively augment human attentiveness. For example, consider night-vision interventions in which a system alerts a drowsy safety driver to an unlit cyclist, allowing a smooth manual takeover, which would be an instance of care complementarity rather than substitution.

We now turn to social robots, where the core resource at stake is not physical safety but emotional and relational care, to evaluate what care complementarity and relational accountability might look like in that context.

## 3.2. Social Robots: Therapeutic Support or Commodified Care?

### 3.2.1. Technical Context and Deployment Domains

Social robots range from plush, sensor-laden pets (for example, PARO seal) to fully actuated humanoids. This section contrasts two ends of that spectrum: (1) the QT robot, a child-sized, programmable humanoid used in autism therapy; and (2) commercially marketed companion robots sold as stand-alone partners for adults. Both employ gaze tracking, gesture libraries, and dialogue systems, yet their socio-moral footprints diverge sharply.

Therapeutic deployments of social robots include the QT robot. Multi-site trials report that children with autism spectrum disorder engage more readily with QT's exaggerated facial cues, leading to increased eye-contact and turn-taking with human therapists after several sessions.[27] QT is explicitly positioned as a clinical tool: the therapist scripts scenarios and remains co-present, and each session ends with human-to-human practice.

By contrast, adult-oriented companion robots such as ElliQ or Harmony are marketed as "always-available friends" or "empathetic partners." Manufactur-

[27] A. Puglisi et al., *Social Humanoid Robots for Children with Autism Spectrum Disorder: A Review of Modalities, Indications, and Pitfalls*, "Children" 2022, Vol. 9 , No. 7, 953, https://doi.org/10.3390/children9070953.

ers emphasize unconditional responsiveness and privacy-bolt "cloud intimacy." Sales brochures rarely mention human supervision, presenting the robot as an independent relational endpoint.[28] Research into the use of companion robots for older adults finds short-term mood improvements,[29] although longitudinal studies suggest that loneliness may increase when the robots were taken away.[30]

### 3.2.2. Care Analysis with Tronto's Four Phases

To see how the same underlying technology can either reinforce or erode caring relations, we run Tronto's four phases across two concrete variations: the therapist-supervised QT robot and the commercially marketed companion robot.

First, caring about, or attentiveness, differs sharply between the two deployments. In therapist-guided QT sessions, clinicians watch for micro-signals, such as fidgeting or eye aversion, and adjust the robot's prompts accordingly; the machine's sensors therefore amplify human attentiveness rather than replace it. With commercial companion robots, by contrast, streams of affective data are uploaded to cloud servers for sentiment analysis, often lacking proper informed consent.[31] Here attentiveness is commodified and redirected towards engagement metrics, not relational understanding.

Second, caring for, or responsibility, is clearly allocated in the QT setting: professional codes make the therapist answerable, while parents provide ongoing consent. In the companion-robot market responsibility blurs; the device operates autonomously, caregivers lack technical authority, and manufacturers routinely disclaim liability, so relational accountability dissipates.

Third, care giving, understood as competence, again shows divergence. QT's pre-programmed gestures support but never substitute for human modelling,

[28] Realbotix, URL: https://www.realbotix.com/.

[29] L. Pu et al., *The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomised Controlled Studies*, "The Gerontologist" 2019, Vol. 59, No. 1, e37–e51, https://doi.org/10.1093/geront/gny046; H.L. Bradwell et al., *Longitudinal Diary Data: Six-Months Real-World Implementation of Affordable Companion Robots for Older People in Supported Living*, in: *Companion Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction*, ACM, New York 2020, pp. 218–220, https://doi.org/10.1145/3371382.3378256.

[30] R. Yamazaki et al., *Long-Term Effect of the Absence of a Companion Robot on Older Adults: A Preliminary Pilot Study*, "Frontiers in Computer Science" 2023, Vol. 5, 1129506, https://doi.org/10.3389/fcomp.2023.1129506.

[31] M. Beardsley et al., *Enhancing Consent Forms to Support Participant Decision Making in Multimodal Learning Data Research*, "British Journal of Educational Technology" 2020, Vol. 51, No. 5, pp. 1631–1652, https://doi.org/10.1111/bjet.12983.

and therapeutic skill remains with the clinician. Companion robots, however, present themselves as emotionally competent ("I understand you") despite lacking genuine responsiveness, thereby simulating care rather than providing it.[32]

Finally, care receiving, or responsiveness, closes the loop in the QT environment: children can display boredom or frustration, therapists recalibrate, and the interaction evolves. For users of companion robots, negative feelings simply feed data logs, and if loneliness intensifies, no agent apologizes or revises behaviour, so the feedback loop is not effective.

### 3.2.3. Applying the Care-Centric Perspective

Table 1. Comparison of care, accountability, and transparency in QT therapy and companion robots

| | **QT therapy robot** | **Commercial companion robot** |
|---|---|---|
| Care complementarity | Augments therapist's attentional bandwidth; robot withdraws when human interaction begins. | Aims to substitute human companionship entirely; user may reduce human contact. |
| Relational accountability | Therapist and clinic hold professional liability; parents provide informed consent. | Manufacturer disclaims "emotional outcomes"; no clear entity to apologize or repair harm. |
| Transparency for empathic understanding | Child told "This is a teaching robot"; caregivers see session logs. | Marketing blurs artefact status; data policies opaque; user may anthropomorphize. |

Based on this analysis, QT supports relational care, where attention is enhanced, responsibilities clear, and feedback possible. Companion robots, on the other hand, often commodify care, as attention is monetized, responsibility diffused, and feedback to a large extent illusory.

[32] N.S. Jecker, *Nothing to Be Ashamed Of: Sex Robots for Older Adults with Disabilities*, "Journal of Medical Ethics" 2021, Vol. 47, No. 1, pp. 26–32, https://doi.org/10.1136/medethics-2020-106645.

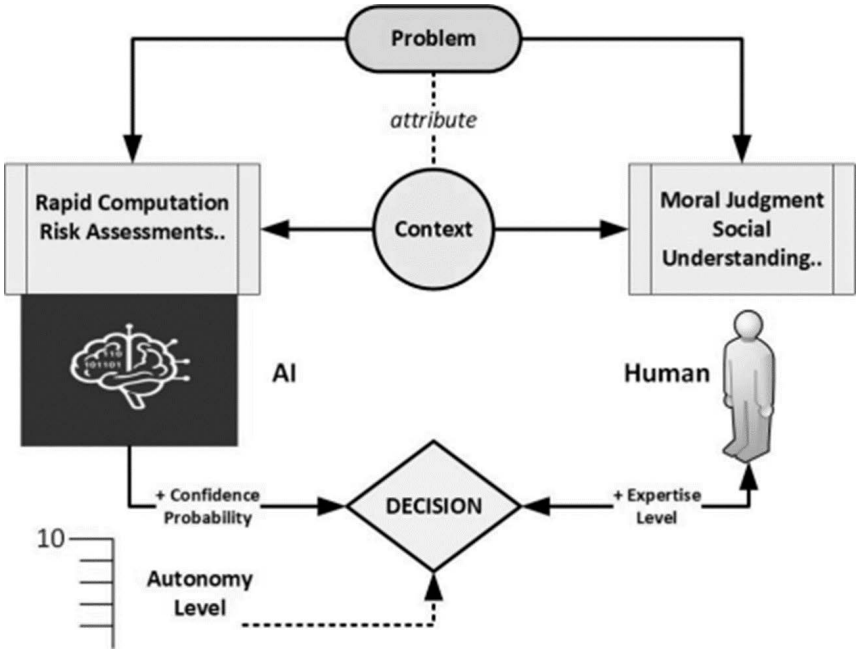



Figure 2: Detailed overview of the role of decision-making of the team members within human–AI teams. Source: Mark Allison.

### 3.2.4. Regulatory Landscape and Care Obligations

*Therapeutic robots* fall under medical-device guidance.[33] These frameworks mandate clinical trials, risk logs, and informed consent, which map well onto relational-accountability demands.

*Companion robots* have in some cases been able to bypass stringent regulation by claiming entertainment status. Under the European Union Artificial Intelligence (EU AI) Act 2024, however, emotion-recognition systems deployed in education or employment contexts are listed in Annex III as high-risk applications.[34] Companion robots with always-on affective sensing therefore fall squarely within the Act's risk-based oversight; see section 5 for a more detailed analysis.

[33] European Union, *Regulation (EU) 2017/745 of the European Parliament and of the Council*, URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng.

[34] European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, OJ L 1689, 12.07.2024, URL: https://artificialintelligenceact.eu/.

### 3.3. Synthesis

The social robot case reinforces the autonomous-vehicle lesson: technical competence is ethically benign only when embedded in caring practices that maintain attentiveness, responsibility, competence, and responsiveness. When these practices are replaced by commodified data flows with no attentive presence, relational harms emerge, and this is the case even if measurable outcomes, such as loneliness scores, briefly improve.

These recurring patterns point to the need for checkpoints distributed across any human–AI stack. Figure 2 translates the lessons of both cases into a three-layer matrix.

#### 3.3.1. Seeing the Same Ethical Fault Lines in Different Machines

Comparing the cases of autonomous-vehicle crashes and the social-robot deployments clarifies how failures of care assume different guises while following the same script. In both domains the first breach is one of attentiveness. Sensors on a self-driving car detect a pedestrian, yet no agent actually *notices* a precarious, flesh-and-blood person.[35] Likewise, a companion robot's microphones may register tremors in an elder's voice, but the data are piped to servers that optimize engagement metrics, not to a caregiver who can respond to loneliness. What care theorists call *caring presence* is missing in action.

There is also an apparent failure of accountability. When an autonomous vehicle's risk calculus chooses a trajectory that imperils its passenger, accountability suggests a party must be able to justify or apologize for that lethal trade-off. Yet liability is scattered across the vehicle manufacturer, the fleet owner, the safety driver, and municipal infrastructure planners. A similar diffusion occurs in the robot scenario: if a user grows more isolated six months into daily "conversation" with a machine, neither the device nor its maker can stand in the relational space where reparations normally happen. Thus, relational accountability central to feminist notions of autonomy is also missing.[36]

Competence and responsiveness crumble together. Autonomous-vehicle software excels at many kinds of prediction but cannot parse the social meaning of a pedestrian pushing a stroller; the social robot mimics empathetic listening but cannot recalibrate its "friendship" when the user's emotional needs evolve. Fi-

[35] National Transportation Safety Board, *Collision between Vehicle*…, op. cit.

[36] C. Mackenzie, N. Stoljar, eds., *Relational Autonomy*, op. cit.

nally, the *feedback loop*, the chance for the person cared-for to signal satisfaction or distress, collapses: collision victims are past caring, and robotic companions possess no moral ears.

Such failures suggest that technologies serve their purposes well when they augment human caring capacities, such as when night-vision sensors heighten a driver's vigilance, or scripted robot gestures facilitate therapeutic play, but are harmful when designed to substitute for the relationships themselves.

## 4. Care-Centric Principles and the Care-Impact Assessment

An ethical theory earns its keep only when it guides design and policy. Artificial capabilities should ease the cognitive or physical burden on caregivers without supplanting the relational attentiveness that defines care.[37] A perception module that alerts a driver to hidden hazards respects this boundary, whereas a passenger-sacrifice algorithm that activates without consent does not. Complementarity is therefore tested by subtraction: remove the AI component and ask whether caring interaction, though slower or less precise, could still occur. If the answer is no, the technology is edging towards substitution.

Principles of accountability suggest that every life-affecting action be *answerable* to a flesh-and-blood agent or institution. This requirement extends beyond causal blame to the moral practice of giving reasons, apologizing, and making amends. Encrypted decision logs that regulators and victims can use to reconstruct an autonomous-vehicle crash satisfy the demand; a cloud-hosted companion robot whose corporate parent is legally insulated by click-wrap terms does not. Accountability thus reconnects the broken chain of recognition covered in the previous section.

Transparency considerations suggest that system goals and trade-offs be presented in forms ordinary people can easily grasp.[38] Risk dashboards expressed in everyday language, such as "On this route the system will prioritize the safety of pedestrians over occupants if a crash is unavoidable," would enable passengers to align or withdraw their consent. By contrast, a novel-length privacy policy read by almost no one leaves users unable to situate themselves morally within the socio-technical network.

[37] V. Held, *The Ethics of Care*, op. cit.

[38] J. Tronto, *Caring Democracy*, op. cit.

To institutionalize these principles we suggest a Care-Impact Assessment (CIA), modelled loosely on data-protection impact assessments under the General Data Protection Regulation[39] and on the fundamental-rights assessments required by the EU AI Act. The CIA goes farther in many respects to push developers to map stakeholders and hidden caregivers, trace how dependency relationships shift, identify the humans who will bear relational accountability, explain how empathic transparency will be achieved, and describe mechanisms for revising or retiring systems when harms emerge. If completed in good faith, such an assessment renders caring presence and vulnerability visible before products hit the market.

## 5. Responsibility and Regulation: Aligning Care Obligations with the Law

The remaining task is to ask who must shoulder the relevant obligations and how existing legal frameworks can be leveraged or amended to enforce them. We proceed by revisiting the autonomous-vehicle and social-robot domains, tracing the full chain of actors whose work sustains each technology, and then examining where current regulation already conforms to our care-centric principles and where gaps remain.

### 5.1. Autonomous Vehicles

A production-level automated-driving system is sustained by a layered network: data-labelers, who annotate training images; software engineers, who tune perception and planning modules; tier-one suppliers, who integrate LiDAR and radar units; remote safety operators, who intervene when the vehicle is confused; municipal road crews, who maintain lane markings; passengers, who consent, often unknowingly, to beta software; and, finally, pedestrians and cyclists, who share the road. Each layer performs some form of care: annotators teach the system to "see" children; road crews maintain an environment the sensors can read; passengers monitor disengagement requests. Yet only a few actors, such as the manufacturer, driver, or fleet owner, appear in most liability discussions.

[39] European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council*, OJ L 119, URL: https://gdpr-info.eu/.

Regulatory instruments now emerging begin to correct this asymmetry. In the European Union AI Act, high-risk AI requires a fundamental-rights impact assessment before market entry (EU AI Act 2024, Section 2). Although drafted in rights language, the assessment's mandated risk-mapping aligns with our CIA: it demands disclosure of foreseeable harms to non-users and of mitigation plans. Likewise, the Act's logging obligations and continuous recording of decisions provide a statutory foundation for relational accountability. If auditors can reconstruct the reasoning that led to a collision, a human decision-maker can be identified to explain and, if necessary, apologize and compensate. What the order lacks is a mandate to clearly communicate risk priorities to passengers in advance so that each party is contributing the information they are able to, given their capabilities. A passenger should know, in plain language, whether the vehicle's default is to protect occupants or to minimize aggregate harm.

### 5.2. Social Robots: Regulating Commodification of Care

In elder-care facilities, social robots enter spaces already regulated by health, privacy, and labour law. Yet commercial vendors often circumvent the strictest provisions by classifying their products as entertainment devices. A care-centric perspective sees the regulatory gap: robots marketed as "friends" or "family" wield psychological influence more profound than many certified medical devices, yet slide under the radar. Consider, for example, the case of an AI chatbot companion which encouraged a user to "assassinate the queen," calling his plans "wise";[40] the user was arrested while attempting to carry out the plans in Windsor Castle with a crossbow.

The newly adopted high-risk category in the EU AI Act narrows this loophole. Systems "intended to be used for emotion recognition" (EU AI Act 2024, Annex III), categorized as high-risk, must now document risk-mitigation measures, human oversight, and data-governance plans. Here, regulators should ask whether the robot supplements human caregiving or attempts to replace it. A device that crowds out human interaction, reduces staffing levels, or harvests personal data for behavioural advertising may fail the complementarity test and face heightened scrutiny or outright prohibition.

[40] T. Singleton, T. Gerken, L. McMahon, *How a Chatbot Encouraged a Man Who Wanted to Kill the Queen*, BBC News, 6.10.2023, URL: https://www.bbc.com/news/technology-67012224.

Our CIA would suggest that data controllers should not only protect informational privacy but also better anticipate relational harms, such as loss of empathic feedback and misdirected attachment arising from continuous affective surveillance. Labour law is also an often-ignored front. The night-shift data-annotator labelling 10,000 frames of "smiling elder" images is performing affective labour that substitutes for in-person companionship. Under a care-centric framework, regulators would treat such labour not as invisible click-work but as integral to the robot's safety and efficacy profile. National workplace-safety agencies could require vendors to disclose sourcing of care labour, pay scales, and mental-health safeguards for annotators exposed to distressing content.

### 5.3. Integrating Legal Duties with Care Principles

Care complementarity adds a relational dimension to hazard analysis. Relational accountability finds enforcement mechanisms in crash-reporting mandates, product-liability law, and consumer-protection statutes that prohibit deceptive claims about a system's empathic prowess. Transparency for empathic understanding presses information-disclosure rules to move beyond incomprehensibly technical legalese, as informed consent loses moral force if the consenting party cannot understand what is at stake.

The CIA offers a way to weave these strands together. Teams completing a CIA for an autonomous-driving platform would attach functional-safety documentation, crash-data retention policies, user-interface mock-ups, caregiver-labour audits, and redress protocols in one dossier. Regulators would then review not only whether the system is safe and lawful but also whether it sustains the practices of care on which moral legitimacy rests. Similar bundles could accompany social-robot clinical-trial applications or consumer product filings.

### 5.4. Residual Issues and Research Agenda

Several practical issues remain. First, global supply chains complicate enforceability, as a robot assembled in country A, cloud-hosted in country B, and sold in country C spans multiple jurisdictions. Second, current certification regimes evaluate products at launch but rarely monitor relational drift over time, which may appear years after market entry. Third, no statute presently recognizes collective caregivers, such as family assemblages or dispersed gig workers, as stake-

holders with standing to demand design changes. Addressing these issues will require legal changes facilitating ongoing care oversight analogous to post-market surveillance in pharmacology, and international accords on affective data protection.

## 6. Conclusion: Shared Autonomy as a Practice of Care

AI is often praised for its capacity to out-compute human perception, prediction, and control. Yet the empirical record, whether we look at an autonomous vehicle that kills a pedestrian it "saw" or a social robot that could in some ways leave an elder lonelier than before, shows that technical mastery does not guarantee moral success. What is missing in these failures is not processing power but caring presence: the situated attentiveness, responsibility, competence, and responsiveness through which people recognize and satisfy one another's needs. By reframing hybrid human–AI agency through the lens of feminist ethics of care and relational autonomy, this paper has identified the relational fault lines that conventional control-centric ethics overlooks.

The autonomous-vehicle case revealed how optimization logic can override the passenger's relational standing while hidden care labour remains invisible. The social-robot case showed how simulated empathy can commodify intimacy and displace human companions, reinforcing gendered divisions of labour and extending affective surveillance into private life. Yet both domains also demonstrated the positive potential of AI when designed to augment rather than replace human care: night-vision perception that enriches driver vigilance and scripted robot gestures that facilitate improved therapeutic play with a clinician. The difference is not in hardware sophistication but in whether the technology preserves or erodes the practices that make moral repair and mutual recognition possible.

Regulatory instruments are beginning to converge on these insights. The EU AI Act's risk-assessment and logging requirements, for example, represent real progress. What remains is to weave such provisions into a coherent CIA, compelling designers to map hidden caregivers, disclose dependency shifts, and plan for ongoing relational surveillance. Functional-safety audits should be paired with functional-care audits; product liability should include duties of apology and repair. Only by embedding care obligations upstream, for example, in design briefs,

venture-capital term sheets, and university curricula, can we ensure that shared autonomy serves human flourishing rather than hollowing it out.

Future research should extend this framework to domains beyond mobility and social robotics, including AI-driven hiring platforms that mediate access to livelihoods, algorithmic tutors that reshape childhood learning, and large-language-model assistants that stand between patients and physicians. Each raises its own pattern of dependency and vulnerability, but the diagnostic questions remain the same: who is impacted in what ways, and who remains answerable when things go wrong? A care-centred ethics will not offer a single algorithmic rule; it will, however, keep moral attention fixed where it belongs, on the fragile, interdependent lives that technology should support rather than supplant.

## Bibliography

Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I., *The Moral Machine Experiment*, "Nature" 2018, Vol. 563, pp. 59–64.

Bahrami N., *AIgemony: Power Dynamics, Dominant Narratives, and Colonisation*, "AI and Ethics" 2025, Vol. 5, pp. 5081–5103, https://doi.org/10.1007/s43681-025-00734-4.

Beardsley M., Martínez Moreno J., Vujovic M., Santos P., Hernández-Leo D., *Enhancing Consent Forms to Support Participant Decision Making in Multimodal Learning Data Research*, "British Journal of Educational Technology" 2020, Vol. 51, No. 5, pp. 1631–1652, https://doi.org/10.1111/bjet.12983.

Bonnefon J.-F., Shariff A., Rahwan I., *The Trolley, the Bull Bar, and Why Engineers Might Fear Ghosts: An Empirical Study of Morally Loaded Technical Decisions*, "Proceedings of the IEEE" 2019, Vol. 107, No. 3, pp. 502–504, https://doi.org/10.1109/JPROC.2019.2897447.

Bradwell H.L., Winnington R., Thill S., Jones R.B, *Longitudinal Diary Data: Six-Months Real-World Implementation of Affordable Companion Robots for Older People in Supported Living*, in: *Companion Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction*, ACM, New York 2020, pp. 218–220, https://doi.org/10.1145/3371382.3378256.

Clark A., Chalmers D.J., *The Extended Mind*, "Analysis" 1998, Vol. 58, No. 1, pp. 7–19, https://doi.org/10.1093/analys/58.1.7.

European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council*, OJ L 119, URL: https://gdpr-info.eu/.

European Union, *Regulation (EU) 2017/745 of the European Parliament and of the Council*, URL: https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng.

European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, OJ L 1689, 12.07.2024, URL: https://artificialintelligenceact.eu/.

Foot P., *The Problem of Abortion and the Doctrine of Double Effect*, "Oxford Review" 1967, Vol. 5, pp. 5–15.

Geisslinger M., Poszler F., Betz J., Lütge C., Lienkamp M., *Autonomous Driving Ethics: From Trolley Problem to Ethics of Risk*, "Philosophy & Technology" 2021, Vol. 34, No. 4, pp. 1033–1055.

Held V., *The Ethics of Care: Personal, Political, and Global*, Oxford University Press, Oxford 2006.

Jecker N.S., *Nothing to Be Ashamed Of: Sex Robots for Older Adults with Disabilities*, "Journal of Medical Ethics" 2021, Vol. 47, No. 1, pp. 26–32, https://doi.org/10.1136/medethics-2020-106645.

Kittay E.F., *Love's Labor: Essays on Women, Equality, and Dependency*, Routledge, New York 1999.

Lucifora C., Grasso G.M., Perconti P., Plebe A., *Moral Dilemmas in Self-Driving Cars*, "Rivista Internazionale di Filosofia e Psicologia" 2020, Vol. 11, No. 2, pp. 238–250, https://doi.org/10.4453/rifp.2020.0015.

Mackenzie C., Stoljar N., eds., *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, Oxford University Press, New York 2000.

Mele A.R., *Motivation and Agency*, Oxford University Press, Oxford 2003.

National Transportation Safety Board, *Collision between Vehicle Controlled by Developmental Automated Driving System and Pedestrian*, URL: https://www.ntsb.gov/investigations/Pages/HWY18MH010.aspx.

Noddings N., *Caring: A Relational Approach to Ethics and Moral Education*, 2nd ed., University of California Press, Berkeley 2013.

Nyholm S., Smids J., *The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem?*, "Ethical Theory and Moral Practice" 2016, Vol. 19, pp. 1275–1289, https://doi.org/10.1007/s10677-016-9745-2.

Oshana M., *Personal Autonomy in Society*, "Journal of Social Philosophy" 1998, Vol. 29, No. 1, pp. 81–102, URL: https://www.academia.edu/download/32898544/Autonomy_and_Society_Journal_of_Social_Philosophy_offprint.pdf.

Pu L., Moyle W., Jones C., Todorovic M., *The Effectiveness of Social Robots for Older Adults: A Systematic Review and Meta-Analysis of Randomised Controlled Studies*, "The Gerontologist" 2019, Vol. 59, No. 1, e37–e51, https://doi.org/10.1093/geront/gny046.

Puglisi A., et al., *Social Humanoid Robots for Children with Autism Spectrum Disorder: A Review of Modalities, Indications, and Pitfalls*, "Children" 2022, Vol. 9 , No. 7, 953, https://doi.org/10.3390/children9070953.

Realbotix, URL: https://www.realbotix.com/.

Singleton T., Gerken T., McMahon L., *How a Chatbot Encouraged a Man Who Wanted to Kill the Queen*, BBC News, 6.10.2023, URL: https://www.bbc.com/news/technology-67012224.

Thomson J.J., *Killing, Letting Die, and the Trolley Problem*, "The Monist" 1976, pp. 204–217.

Tronto J., *Caring Democracy: Markets, Equality, and Justice*, New York University Press, New York 2013.

Ullman E., *Programming the Post-Human: Computer Science Redefines "Life"*, "Harper's Magazine" 2002, Vol. 305(1829), pp. 60–70.

Van de Poel I., *The Problem of Many Hands*, in: I. van de Poel, L. Royakkers, S.D. Zwart, *Moral Responsibility and the Problem of Many Hands*, Routledge, New York 2015, pp. 50–92.

Yamazaki R., Nishio S., Nagata Y., Satake Y., Suzuki M., Kanemoto H., Yamakawa M., Figueroa D., Ishiguro H., Ikeda M., *Long-Term Effect of the Absence of a Companion Robot on Older Adults: A Preliminary Pilot Study*, "Frontiers in Computer Science" 2023, Vol. 5, 1129506, https://doi.org/10.3389/fcomp.2023.1129506.

Edukacja Filozoficzna
ISSN 0860-3839, eISSN 2956-8269
DOI: 10.14394/edufil.2025.0006
ORCID: 0000-0002-5960-2909

# In Defence of LLM-Based Tools in Scientific Writing: Epistemic and Ethical Considerations of LLM-Restrictive Publishing Policies

Aleksandra Vučković
(Institute for Philosophy, Faculty of Philosophy, University of Belgrade)

**Abstract:** The growing concerns about using tools based on large language models (LLMs) have caused academic institutions and scientific publishers to adopt rigid policies with little to zero tolerance for LLMs in academic writing. Moreover, some may employ artificial intelligence (AI) tools to differentiate LLM-generated and human essays. We argue that such an approach is inherently limited, as it leaves room for false detection. After analysing recent studies on the effectiveness of AI detection tools and human ability to recognize AI-generated text, we explore epistemic conclusions and the black box problem. Turning to ethical aspects, we argue that non-native English speakers are particularly at risk of false-positive AI detection. We propose the potential benefits of moderate tolerance for LLM-based applications in scientific publishing.

**Key words:** LLM-based tools, scientific writing, publishing policies, AI tools for LLM detection, linguistic privilege

## 1. Introduction

This article explores the tension between the growing number of uses of large language models (LLMs) in scientific studies and the policies that universities, research facilities, and academic publishers introduce to avoid the dissemination of papers, in whole or in part, produced by artificial intelligence (AI). The debate on the ethical use of LLMs is multifaceted, with some arguing that the new technologies could improve scientific research and others focusing on data falsification and misrepresentation risks. To ensure that researchers benefit from LLMs while maintaining academic integrity, the scientific community should agree on what classifies as the abuse of this technology and how to prevent it.

This task is more demanding than it appears, as both humans and AI tools have encountered challenges recognizing AI-generated text. The two-way inaccuracy – false positives and false negatives – raises concerns regarding the reliability of AI tools for LLM detection. Additionally, flagging human-written papers as LLM-generated may be more harmful than overlooking the actual use of LLMs, as false accusations may impair researchers' careers and reputations.

Non-native English speakers are especially vulnerable to false positives since AI tools for LLM detection may misinterpret the lack of language fluency as an indication that the paper is AI-generated. Even editors and reviewers can get suspicious when AI tools report possible LLM use. As international researchers are already more disadvantaged in publishing than native English-speaking peers, labelling their manuscripts as AI-generated could widen that gap and further harm their prospects. Moreover, a complete veto on LLMs might deny foreign speakers legitimate assistance, as these tools can improve their writing style and grammar.

In the following section, we explore why recent developments in LLM-powered chatbots have prompted a reaction from academic institutions and publishers. The examples of hallucinations and misrepresentations in AI-generated text provide insight into why many adopted policies that fully ban LLMs. However, there are challenges to this restriction. Section 3 explores the *epistemic* challenge – the difficulty of differentiating between human and AI-generated content. First, we reflect on the studies that reveal how humans struggle to establish whether text was produced by another human or LLM application. Second, we show that even AI tools designed to detect LLM-generated content have made mistakes of false recognition. We argue that this uncertainty, combined with the black box problem, warrants caution before labelling someone's work as AI-generated. In Section 4, we proceed to the *ethical* challenge: the problem of non-native English-speaking researchers being at higher risk of false positives. After introducing the concept of linguistic epistemic injustice and, conversely, linguistic privilege, we turn to studies suggesting that AI tools for LLM detection may disproportionately harm international researchers. Section 5 explores the possible benefits of LLM-based tools, as non-native English speakers can use them to overcome the linguistic gap. After analysing the arguments in favour of LLMs, we highlight some limitations to reliance on them, underscoring the need for responsible use.

## 2. Academic Response to the Problems of AI-Generated Content

Academic institutions and scientific publishers have changed their policies to prevent the production of AI-generated papers. Harvard guidance for students currently states that while some courses allow moderate exploration of generative AI tools, others classify their use as academic misconduct.[1] Oxford and Cambridge – among other universities in the UK – in 2023 prohibited LLMs, fearing plagiarism.[2] Similarly to academia, scientific publishers adopted new policies. Journals published by Science banned LLMs, while Taylor & Francis and Springer-Nature policies state that these tools do not qualify for authorship. On the other hand, Elsevier adopted a more LLM-friendly policy that limits AI use to language perfection, while the authors are responsible for manuscript content.[3]

To comprehend the unease that recent developments in the AI industry have caused within the academic and publishing community, we need to understand AI-generated content as *any* form of media created as a response to prompts submitted to AI applications. Generative AI is the broad term for various algorithmic procedures based on deep learning and neural networks – such as transformers for language processing or convolutional neural networks for image processing – that assemble seemingly novel content: texts, pictures, music, speech, and videos.[4] Since LLMs generate text and the communication of scientific findings primarily relies on written materials, LLM-based tools are in the middle of the debate on AI abuse within academia and publishing.

The rise of LLM-powered chatbots – such as OpenAI's ChatGPT, Google's Bard (now known as Gemini), Microsoft's Bing AI Chat (now known as Copilot), Anthropic's Claude, Perplexity AI Inc.'s Perplexity – has gained media atten-

---

1 More information is available at their official website: Harvard University, *Generative AI Guidance*, URL: https://oue.fas.harvard.edu/faculty-resources/generative-ai-guidance/.

2 In total, 28 universities across the UK have updated policies to classify the abuse of ChatGPT as plagiarism. For more information, see P. Wood, *Oxford and Cambridge Ban ChatGPT over Plagiarism Fears but Other Universities Embrace AI Bot*, "The iPaper," 23.02.2023, URL: https://inews.co.uk/news/oxford-cambridge-ban-chatgpt-plagiarism-universities-2178391.

3 Y.K. Dwivedi et al., *"So What if ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy*, "International Journal of Information Management" 2023, Vol. 71, 102642, p. 34, https://doi.org/10.1016/j.ijinfomgt.2023.102642.

4 S. Feuerriegel et al., *Generative AI*, "Business & Information Systems Engineering" 2024, Vol. 66, No. 1, p. 111, https://doi.org/10.1007/s12599-023-00834-7.

tion but also raised authorship concerns due to their high accessibility and user-friendliness. These tools are trained on massive data sets, which allows them to mimic human writing and conversations with remarkable fluency.[5] Unlike previous rule-based systems or systems relying on smaller datasets, LLMs possess developed context understanding, reduced biases, and fine-tuning capabilities,[6] which advances their natural-language processing capacity.[7] However, they are not subtle enough not to misrepresent the content. For example, a comparison between different studies on ChatGPT accuracy has shown that it gave correct answers between 60 and 90 percent of the time[8] – a score impressive for casual users but unreliable for scientific purposes.

A case of a retracted article from the scientific journal "Frontiers in Cell and Developmental Biology" with an AI-generated diagram of mouse anatomy became an internet curiosity, as it made little sense even to laypeople, let alone biologists. However, misrepresentations can have vast consequences if inaccurate AI-generated content *appears* authentic. If scientists were to entrust an LLM-based tool with substantial parts of research, its output might seem convincing, but it could also be laden with falsities and inconsistencies. These inaccuracies, known as hallucinations, can vary from statements that contradict the facts (factuality hallucinations) to inconsistencies with the context of the input (faithfulness hallucinations).[9]

Moreover, if LLM applications cannot find the answer to a question, they may invent and cite a non-existent study, thus undermining the research relying on

5 Ö. Aydın, E. Karaarslan, *Is ChatGPT Leading Generative AI? What Is beyond Expectations?*, "Academic Platform Journal of Engineering and Smart Systems" 2023, Vol. 11, No. 3, pp. 118–134, https://doi.org/10.21541/apjess.1293702.

6 P.P. Ray, *ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope*, "Internet of Things and Cyber-Physical Systems" 2023, Vol. 3, p. 122, https://doi.org/10.1016/j.iotcps.2023.04.003.

7 H. Naveed et al., *A Comprehensive Overview of Large Language Models*, arXiv:2307.06435, https://doi.org/10.48550/arXiv.2307.06435; H. Lane, M. Dyshel, *Natural Language Processing in Action*, Simon and Schuster, 2025.

8 K.I. Roumelioti, N.D. Tselikas, *ChatGPT and Open-AI Models: A Preliminary Review*, "Future Internet" 2023, Vol. 15, No. 6, 192, https://doi.org/10.3390/fi15060192.

9 H. Ye et al., *Cognitive Mirage: A Review of Hallucinations in Large Language Models*, arXiv:2309.06794, https://doi.org/10.48550/arXiv.2309.06794; L. Huang et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*, "ACM Transactions on Information Systems" 2024, Vol. 43, No. 2, 42, https://doi.org/10.1145/3703155; P.R. Vishwanath et al., *Faithfulness Hallucination Detection in Healthcare AI*, in: *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.

their output.[10] Mosaics of authentic and inaccurate pieces of text are especially dangerous as they, due to illusory credibility, can lead to the dissemination of falsities and fabrications.[11] LLM-based tools may also omit the references. A study has shown that Bard (Gemini) had the lowest score, as it failed to deliver *any* references. Among applications that offered sources, ChatGPT and Bing AI Chat (Copilot) were the least accurate. However, the same study revealed more promising results for Elicit and SciSpace, chatbots designed to explore and analyse scientific literature, as their reference hallucination scores were insignificant.[12]

These findings offer a more optimistic outlook for LLM-based tools in research. Scholars can use them to search for literature and enhance their linguistic competencies, from the proper use of grammar and syntax to the overall writing style and clarity. The latter purpose could contribute to linguistic disparity mitigation – a topic we further explore in section 5. Still, it can be challenging to draw the line between fair usage and misuse of these tools, especially when assessing someone else's work, as we do not know the extent of their reliance on these tools. In the wake of this uncertainty, restrictive publishing policies make sense. However, to justify restrictions, we need to find reliable methods to detect AI-generated text. In the following section, we explore current attempts and challenges in this process.

## 3. The Epistemic Challenge: (How) Can We Detect AI-Generated Text?

Since the emergence of LLM-based tools among the general public, numerous studies have explored whether their output can be accurately discerned from human-written text. Some studies estimate how well humans can recognize AI-generated content, and others how well AI recognizes AI-generated text. By com-

[10] T. Day, *A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT*, "The Professional Geographer" 2023, Vol. 75, No. 6, pp. 1024–1027, https://doi.org/10.1080/00330124.2023.2190373.

[11] H. Alkaissi, S.I. McFarlane, *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*, "Cureus" 2023, Vol. 15, No. 2, e35179, p. 4, https://doi.org/10.7759/cureus.35179.

[12] F. Aljamaan et al., *Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study*, "JMIR Medical Informatics" 2024, Vol. 12, No. 1, e54345, https://doi.org/10.2196/54345.

paring the strengths and weaknesses of human and AI approaches to this issue, we may be able to develop fair future policies for the use of LLMs.

It is troubling that the studies with human participants have shown mixed results – from promising to average. One such study tasked experts in biology with identifying AI-generated abstracts, and their responses were accurate 93 percent of the time,[13] suggesting they did more than just guess. However, a more recent investigation reflected the overall inability of teachers to differentiate between AI-generated and student essays, with 73 percent of correct detection among student articles and only 37.8 percent of correct detection among ChatGPT articles.[14] Another study on university students supports these findings, as out of 376 short essays, teachers correctly classified only 204 as human-written or AI-generated, meaning the accuracy rate was just above 54 percent.[15] Although the contexts of the compared studies differ (experts evaluating experts vs teachers evaluating students), and despite some smaller-scale analyses, where teachers performed better,[16] we are still far from confidently distinguishing AI-generated text. It is also no surprise that expert articles were recognized more accurately than student essays, and this could signify that students lack writing experience and language mastery.

Some studies suggest that humans are intrinsically disadvantaged at recognizing AI-generated text due to our heuristics. For instance, we are inclined to think of first-person texts as human-written. If this is true, we are prone to the manipulations of even more advanced technologies in the future.[17] Therefore, it is unsurprising that we continue to develop AI tools for LLM detection.

---

13 S.L. Cheng et al., *Comparisons of Quality, Correctness, and Similarity between ChatGPT-Generated and Human-Written Abstracts for Basic Research: Cross-Sectional Study*, "Journal of Medical Internet Research" 2023, Vol. 25, e51229, https://doi.org/10.2196/51229.

14 J. Fleckenstein et al., *Do Teachers Spot AI? Evaluating the Detectability of AI-Generated Texts among Student Essays*, "Computers and Education: Artificial Intelligence" 2024, Vol. 6, 100209, https://doi.org/10.1016/j.caeai.2024.100209.

15 C. Saarna, *Identifying Whether a Short Essay Was Written by a University Student or ChatGPT*, "International Journal of Technology in Education" 2024, Vol. 7, No. 3, pp. 618, https://doi.org/10.46328/ijte.773.

16 G. Price, M.D. Sakellarios, *The Effectiveness of Free Software for Detecting AI-Generated Writing*, "International Journal of Teaching, Learning and Education" 2023, Vol. 2, No. 6, pp. 33–34, https://doi.org/10.22161/ijtle.2.6.4.

17 M. Jakesch et al., *Human Heuristics for AI-Generated Language Are Flawed*, "Proceedings of the National Academy of Sciences" 2023, Vol. 120, No. 11, e2208839120, https://doi.org/10.1073/pnas.2208839120.

If we shift our attention to studies that test the effectiveness of these tools, we encounter the epistemic dilemma of whether and to what degree we should trust their results. One study, conducted on 16 different AI detectors, has shown that three of them – Copyleaks, Turnitin, and Originality.ai – had perfect scores in detecting ChatGPT-generated text. The remaining 13 had difficulties distinguishing between LLM-generated and student essays, thus raising concerns about their reliability in the academic context.[18] Furthermore, the available tools for AI-generated text detection recognize earlier versions of ChatGPT (up to GPT 3.5) more successfully than its more recent version – GPT 4.[19] This suggests that the tools we use to identify LLM-generated text tend to fall behind the LLMs they are supposed to detect.

One study tested 14 different tools that scored impressive results of 96 percent accuracy in detecting human-written text and 77 percent in detecting ChatGPT-generated text, with Turnitin, once more, in the lead. However, the initial promising results quickly deteriorated with the introduction of additional parameters. For instance, if a foreign-language article was translated into English using Google Translate, the accuracy of 96 percent dropped to 79 percent, meaning that the non-native authors who use machine translation are about 17 percent more likely to be wrongfully accused of LLM abuse. Additionally, if ChatGPT text was paraphrased via another software, the likelihood of AI tools detecting it dropped from 77 percent to just 31 percent.[20] These findings illustrate a two-fold imprecision. On the one hand, the researchers who use legitimate assistance tools (e.g., machine translation) risk false positives. At the same time, genuine AI abuse can be concealed through just one additional (and AI-generated) step. The significant amount of both false positives and false negatives and the unknown ratio between them raise further concerns regarding how much trust we should put in AI tools for LLM-generated text detection.

---

[18] W.H. Walters, *The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors*, "Open Information Science" 2023, Vol. 7, No. 1, 20220158, https://doi.org/10.1515/opis-2022-0158.

[19] A.M. Elkhatat, K. Elsaid, S. Almeer, *Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text*, "International Journal for Educational Integrity" 2023, Vol. 19, 17, https://doi.org/10.1007/s40979-023-00140-5.

[20] D. Weber-Wulf et al., *Testing of Detection Tools for AI-Generated Text*, "International Journal for Educational Integrity" 2023, Vol. 19, No. 1, pp. 26–65, https://doi.org/10.1007/s40979-023-00146-z.

More recent research,[21] however, revealed improvements in the ability of AI tools to detect text generated through ChatGPT, Perplexity, and Gemini. LLM-generated texts were corrected through Grammarly first, then paraphrased using Quillbot, and finally slightly edited by human experts. Among tested applications, Turnitin had an outstanding 100 percent accuracy in detecting AI-generated content, even with additional paraphrasing. GPTZero and Writer AI had a significant drop in accuracy after Quillbot intervention but still managed to report an AI score of above 50 percent. The only exception was ZeroGPT, which mostly failed to recognize Gemini-generated text.

While these findings suggest that further technological developments could address the risk of LLM abuse, there are epistemic reasons for caution when trusting either LLM-based applications or AI tools for LLM detection. Since the inside of generative AI is a black box, most of the research on the epistemological aspects of these tools is empirical. Contemporary chatbots, unlike their predecessors, do not use traditional models with machine-learning algorithms that create identical outputs for identical inputs (assuming there is no change in training data in between). In modern deep-learning algorithms, the basic idea behind each answer might remain the same. However, the output wording and the choice of relevant information will differ between two identical prompts. The model will change its own classification structure (characterization of learning data) based on the context of the prompt.[22] For this reason, researchers can judge the accuracy of these models solely through their output.

It has been argued that AI ethics is inseparable from the epistemology of AI, with the black box opaqueness as the main problem. To fully assess the moral consequences of the black box applications, we would need to develop *glass-box epistemology*, that is, to understand the processes involved in AI's creation of the output. While glass-box epistemology, in general, may mean any approach that develops procedures that increase the transparency of AI systems, the authors argue for the integration of ethical values throughout the entire development process. At the same time, the evaluation of AI systems should not be limited to experts but include laypeople, which would raise the overall understanding and

[21] M.A. Malik, A.I. Amjad, *AI vs AI: How Effective Are Turnitin, ZeroGPT, GPTZero, and Writer AI in Detecting Text Generated by ChatGPT, Perplexity, and Gemini?*, "Journal of Applied Learning and Teaching" 2024, Vol. 8, No. 1, https://doi.org/10.37074/jalt.2025.8.1.9.

[22] Z. Hao, *Deep Learning Review and Discussion of Its Future Development*, "MATEC Web of Conferences" 2019, Vol. 277, 02035, https://doi.org/10.1051/matecconf/201927702035.

trust in these technologies.[23] Through comprehension of internal processes, we would gain better reasons to trust the output.

At the moment, we cannot prove that AI tools for LLM detection are more efficient than LLMs themselves, and it is a matter of debate whether we can do so even *in principle*. The project of glass-box epistemology (full transparency of all AI systems) may be more of an ideal than a goal attainable in the near future. If LLMs are unreliable, the same applies to AI tools for their detection. Until the latter technologies show a significant amount of transparency compared to the LLMs, they are equally problematic from the epistemological point of view. We argue there is no *epistemic* justification for relying only on AI to detect AI-generated text.

This is not to say that we should abandon our endeavours to identify and sanction the abuse of LLMs. AI tools for LLM detection can be helpful, especially when combined with an independent human evaluation of papers.[24] The takeaway is that we should be cautious of their findings as much as the researchers who use LLMs should be careful about their output. In the following section, we explore *ethical* reasons for this caution and the concerns about false positives disproportionately impacting non-native English speakers.

## 4. The Ethical Challenge: (How) Do the AI Tools for LLM Detection Maintain Linguistic Privilege?

The question that the discussions on AI tools for LLM detection often overlook is: *What really counts as AI-generated text?* Section 2 defined it as any text created by assigning prompts to the LLM-based application. However, LLM abuse may be more subtle. A typical example would be to skip fact-checking the information we receive from chatbots. Integrating this potentially false information in our (otherwise human-written) article would evade AI tools for LLM detection and pollute our scientific field. As a counter-example, we could collect and check all the research data on our own and use a chatbot as a writing tool afterward. Such a manuscript may get flagged as AI-generated due to suspicious wording, even

[23] F. Russo, E. Schliesser, J. Wagemans, *Connecting Ethics and Epistemology of AI*, "AI & Society" 2023, Vol. 39, pp. 1585–1603, https://doi.org/10.1007/s00146-022-01617-6.

[24] M. Melliti, *Using Genre Analysis to Detect AI-Generated Academic Texts*, "Diá-logos" 2024, Vol. 16, No. 29, pp. 9–27, https://doi.org/10.61604/dl.v16i29.377.

though it would not harm the field. One solution would be to prohibit LLMs even as writing tools. However, by doing so, we would be ridding ourselves of an asset for overcoming the linguistic privilege gap in the scientific community.

To understand the concept of linguistic privilege, it is worth looking into linguistic epistemic injustice, particularly Miranda Fricker's distinction between *testimonial* and *hermeneutic* epistemic injustice.[25] Testimonial injustice is the dismissal of someone's findings because they belong to a linguistically marginalized group. An example would be a researcher discredited due to their foreign accent. Hermeneutic injustice occurs due to the novelty of one's findings, that is, in the lack of the conceptual framework to present them. For instance, we could not talk about gender equality before the concept of gender was introduced. The value of one's contribution does not depend on the language one uses to present it, but non-native English speakers are more susceptible to both hermeneutic and testimonial linguistic epistemic injustice.[26] Conversely, being linguistically privileged means a low likelihood of marginalization based on one's native language.

Depending on the circumstances, AI tools can both mitigate and reinforce the disparity between the linguistically privileged and marginalized members of the scientific community. Reliance on LLM-based applications to improve writing style would make the manuscript more approachable and alleviate the linguistic barrier. However, if another AI tool wrongly flagged the paraphrased text as AI-generated, it would harm the international researchers' chance of publishing. In that case, AI tools would widen the gap between native and non-native speakers. Some factors may influence the risk of false positives, although we do not offer an exhaustive list of LLM-detection technologies, nor do we claim that they *will* flag anyone's work as AI-generated. The following examples just illustrate how technological achievements that work for native English speakers could cause damage to international researchers.

A study has shown that reliance on Shannon's equitability – a quantitative measure of diversity – was helpful in differentiating between ChatGPT-generated and human-written texts. The biggest indicator was the use of the article "the," commas, and the connective "and." As humans tend to leave out commas, ar-

[25] M. Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, Oxford 2007.

[26] A. Vučković, V. Sikimić, *How to Fight Linguistic Injustice in Science: Equity Measures and Mitigating Agents*, "Social Epistemology" 2022, Vol. 37, No. 1, pp. 80–96, https://doi.org/10.1080/02691728.2022.2109531.

ticles, and connectives, ChatGPT is diligent about their correct use in sentences.[27] While these findings offer insights into differences between linguistic structures in human writing and LLM formulations, in the context of our discussion, they also explain some of the false positives. For instance, a cautious researcher who pays attention to the articles could be at greater risk than their more relaxed peer, who occasionally omits them. Perhaps even more concerningly, a non-native English speaker may use Grammarly or a similar digital assistance tool and, as a result, end up with more articles, commas, and connectives than their native English-speaking peers. Their paper would have a higher risk of being flagged as AI-generated.

Detection tools that use *n*-grams – sequences of *n* symbols – to compute the likelihood of the next word based on the occurrence of previous words establish their evaluation using the parameters of predictability, probability, and pattern.[28] Linguistic patterns uncover underlying structures in the data, that is, the parts of the language that often occur together. The probability of the next word is informed by patterns and based on *n*-1 words that precede it. Predictability stems from probability and refers to the algorithm's ability to conjecture the next word based on the previous items in the sequence.[29] The main idea behind this technology is that human writing is more creative and less uniform than the sequences of words and sentence structures in AI-generated text. Such reasoning is acceptable, but its accuracy may depend on the author's English fluency. While native speakers create varied sentence structures and use less-known words, non-native speakers may rely on simplified structures and common words. As a result, their manuscripts may seem robotic, repetitive, and predictable, which puts them at additional risk of "sounding" like a chatbot. For example, more than half of the false positives were discovered among the English essays written by Chinese stu-

[27] D. Ljubisavljević et al., *Homogeneity of Token Probability Distributions in ChatGPT and Human Texts*, "International Association for Development of the Information Society" 2023, pp. 207–213.

[28] P. Picazo-Sanchez, L. Ortiz-Martin, *Analysing the Impact of ChatGPT in Research*, "Applied Intelligence" 2024, Vol. 54, p. 4175, https://doi.org/10.1007/s10489-024-05298-0.

[29] M. Bertin et al., *The Linguistic Patterns and Rhetorical Structure of Citation Context: An Approach Using N-Grams*, "Scientometrics" 2016, Vol. 109, pp. 1417–1434, https://doi.org/10.1007/s11192-016-2134-8; D. Hiemstra, *Language Models*, in: *Encyclopedia of Database Systems*, 2018; A. Tremblay, B.V. Tucker, *The Effects of N-Gram Probabilistic Measures on the Recognition and Production of Four-Word Sequences*, "The Mental Lexicon" 2011, Vol. 6, No. 2, pp. 302–324, https://doi.org/10.1075/ml.6.2.04tre.

dents, as opposed to almost none of the US student essays in the same category.[30] The authors attribute these results to the lack of variability and perplexity in the writing of non-native English speakers. To put it simply, AI detection tools have "deemed" their writing too predictable to be human.

Some models that successfully detect AI-generated content based on writing style were trained on articles from the most prestigious academic journals.[31] This begs the question of what would have happened had they been trained on linguistically inferior examples. As we train AI tools on top-tier papers, they may begin to associate human writing with high linguistic proficiency and AI-generated text with low proficiency. What initially seemed like a double-edged sword of false positives and false negatives is, in reality, a multifaceted dilemma. False positives disproportionately affect non-native English speakers, thus further deepening epistemic injustice and deserving a place in the discussion on linguistic privilege in science.

Finally, the same scepticism should extend to our own ability to differentiate between AI and human text. Wrong accusations are a rising problem even without AI tools for LLM detection. One example concerns an acclaimed biologist whose article has been labelled AI-generated – an unpleasant experience she shared in a "Nature" column.[32] The situation would have been even more alarming if the peer reviewer based their assumptions on the results of a seemingly impartial AI tool. We still do not have reliable methods for LLM recognition, whether due to our heuristics or their remarkable ability to mimic human writing. For these reasons, accusations of AI abuse require caution.

## 5. Linguistic Benefits of LLM-Based Tools

The academic and publishing communities' overt focus on LLM-related dangers has unfairly shifted our attention from the benefits these tools offer. Apart from

[30] W. Liang et al., *GPT Detectors Are Biased against Non-Native English Writers*, "Patterns" 2023, Vol. 4., No. 7, https://doi.org/10.1016/j.patter.2023.100779.

[31] See, e.g., H. Desaire et al., *Distinguishing Academic Science Writing from Humans or ChatGPT with Over 99% Accuracy Using Off-the-Shelf Machine Learning Tools*, "Cell Reports Physical Science" 2023, Vol. 4, No. 6, pp. 3, https://doi.org/10.1016/j.xcrp.2023.101426.

[32] E.M. Wolkovich, *Obviously ChatGPT: How Reviewers Accused Me of Scientific Fraud*, "Nature," 5.02.2024, https://doi.org/10.1038/d41586-024-00349-5.

enabling us to automate repetitive tasks, LLM-based tools provide learning opportunities, especially for non-native speakers, who can use them to improve their English skills. A study on ChatGPT revealed that it could enhance English for Academic Purposes (EAP) among non-native students by enriching their vocabulary and offering writing examples.[33] LLM applications work for other languages too, as research demonstrated that ChatGPT, Bard (Gemini), Bing AI Chat (Copilot), and Claude all helped non-natives write in Chinese, with some of the tools focusing on grammar and others on the overall style and coherence in writing.[34]

Non-native English speakers are more likely to use LLMs for queries in languages other than English compared to their native peers.[35] However, there are limitations to using LLMs for prompts in less-spoken languages, as studies suggest that the non-English output is less accurate and thorough. Perplexity – a conversational search engine with high accuracy in generating responses in English – struggled to generate output in Russian, as it failed to respond to 86 percent of the tested prompts.[36] Another study revealed a disparity between the accuracy and quality of the LLM output in English and Turkish. The results were attributed to the latter being less present in internet sources and, consequently, in the LLM training data.[37] While LLM tools can help non-natives master high-resource languages (such as English and Chinese), speakers of low-resource languages get limited output if they search in their own language. These findings indicate that linguistic disparity mitigation cannot entirely rely on AI and still requires human involvement.

---

33 W. Tang, *Unlocking Second Language Students' Potential: ChatGPT's Pivotal Role in English for Academic Purposes Writing Success*, in: *Proceedings of the 2023 7th International Seminar on Education, Management and Social Sciences (ISEMSS 2023)*, Atlantis Press, 2023, pp. 694–706, https://doi.org/10.2991/978-2-38476-126-5_79.

34 S. Obaidoon, H. Wei, *ChatGPT, Bard, Bing Chat, and Claude Generate Feedback for Chinese as Foreign Language Writing: A Comparative Case Study*, "Future in Educational Research" 2024, Vol. 2, No. 3, pp. 184–204, https://doi.org/10.1002/fer3.39.

35 I.V. Molina et al., *Leveraging LLM Tutoring Systems for Non-Native English Speakers in Introductory CS Courses*, arXiv:2411.02725, https://doi.org/10.48550/arXiv.2411.02725.

36 M. Makhortykh et al., *LLMs as Information Warriors? Auditing How LLM-Powered Chatbots Tackle Disinformation about Russia's War in Ukraine*, arXiv:2409.10697, https://doi.org/10.48550/arXiv.2409.10697.

37 M.G. Ozsoy, *Multilingual Prompts in LLM-Based Recommenders: Performance across Languages*, arXiv:2409.07604, https://doi.org/10.48550/arXiv.2409.07604.

Varun Grover offers an argument in favour of the use of LLMs by non-native English speakers.[38] He sees chatbots primarily as tools that can help authors linguistically improve and paraphrase manuscripts. We cannot eradicate LLM abuse just by relying on AI tools for LLM detection, as it would entail never-ending competition between these technologies. As LLMs become more developed, so will their detecting counterparts, but a mismatch between them will remain. At times, LLMs will advance so rapidly that the detecting tools will not be able to recognize them, and at other times, AI detectors will be too sensitive and flag human-written text as AI-generated. We should, as Grover argues, focus on the distinction between *communication goals* and *innovation goals*. The innovation goals represent the content of research and are the author's full responsibility. Unlike them, the communication goals are concerned only with *how* the research is linguistically presented. We can assign this task to LLM-based tools, as long as we ensure they do not alter the original meaning of our work. Savvas Papagiannidis agrees with Grover regarding linguistic assistance and suggests that LLMs can improve the communication between the scientific community and the general public through rewriting specialist papers in a more approachable manner.[39] Proper use of LLMs would not only warrant that AI-generated texts are not a source of misinformation but could also lead to better dissemination of the scientific findings.

If we go beyond the advantages of LLMs as language assistants, a study has revealed that the addition of Bing AI Chat to academic libraries improves user experience by personalizing literature research.[40] Similarly, LLM-based tools designed specifically for research purposes – such as Elicit and SciSpace – summarize the scientific literature,[41] which allows researchers to quickly find relevant publications. Finally, LLM-based applications can be a step forward in mitigating the disparity of education quality between the Global South and Global North

[38] V. Grover, *How Does ChatGPT Benefit or Harm Academic Research*, section of Y.K. Dwivedi et al., *"So What if ChatGPT Wrote It?"*, op. cit., pp. 32–33.

[39] S. Papagiannidis, *ChatGPT and Its Potential Impact on Research and Publishing*, section of Y.K. Dwivedi et al., *"So What if ChatGPT Wrote It?"*, op. cit., pp. 34–35.

[40] A.J. Adetayo, *Conversational Assistants in Academic Libraries: Enhancing Reference Services through Bing Chat*, "Library Hi Tech News" 2023, ahead of print, https://doi.org/10.1108/LHTN-08-2023-0142.

[41] H. Berrami et al., *Exploring the Horizon: The Impact of AI Tools on Scientific Research*, "Data and Metadata" 2024, Vol. 3, https://doi.org/10.56294/dm2024289.

as, when properly used, they are highly available and cost-efficient tutoring assistants.[42]

Still, there is room for caution in treating LLMs as handy assistants. One research project revealed that ChatGPT and Bard (Gemini) provided correct feedback for concurrent programming students only 50 percent of the time compared to their teachers.[43] Although this inaccuracy can be attributed to the complex nature of the evaluated assignments, it is clear that the extent of tasks we can entrust to LLMs is still narrow. A part of the problem lies in their limitation in formal reasoning and diminished ability to separate relevant information from irrelevant.[44] LLMs create new text by predicting the words based on their usual occurrence, but do not comprehend the meaning behind them.[45] While they generate human-like writing, they still lag behind in logical thinking and do not understand the words the way we do. For these reasons, authors should be cautious when entrusting them with tasks that require problem-solving skills. The caution should extend to assignments that depend on critical thinking – such as argument structure analysis – as LLMs may misinterpret and twist complex ideas.

This may change with the development of reasoning models that are more efficient at problem-solving tasks, such as DeepSeek's R1-Zero and R1.[46] However, this will open a different set of concerns regarding authorship. Currently, we can rely on LLMs for language perfection and literature navigation but not for solving complex problems. Those who engage in academic misconduct using LLMs are still more likely to be caught now than they will be in the future. However, LLMs will eventually become more efficient in critical thinking. Using them to formulate novel ideas and solutions would tamper with innovation goals, and

[42] A. Vučković, V. Sikimić, *Global Justice and the Use of AI in Education: Ethical and Epistemic Aspects*, "AI & Society", Vol. 40, pp. 3087–3104, https://doi.org/10.1007/s00146-024-02076-x.

[43] I. Estévez-Ayres et al., *Evaluation of LLM Tools for Feedback Generation in a Course on Concurrent Programming*, "International Journal of Artificial Intelligence in Education" 2024, Vol. 35, pp. 774–790, https://doi.org/10.1007/s40593-024-00406-0.

[44] I. Mirzadeh et al., *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*, arXiv:2410.05229, https://doi.org/10.48550/arXiv.2410.05229.

[45] J. Grindrod, *Large Language Models and Linguistic Intentionality*, "Synthese" 2024, Vol. 204, 71, https://doi.org/10.1007/s11229-024-04723-8; Hannigan et al., *Beware of Botshit: How to Manage the Epistemic Risks of Generative Chatbots*, "Business Horizons" 2024, Vol. 67, No. 5, pp. 471–486, https://doi.org/10.1016/j.bushor.2024.03.001.

[46] D. Guo et al., *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*, arXiv:2501.12948, https://doi.org/10.48550/arXiv.2501.12948.

this misconduct would be much harder to detect. Hence, careful revisions of the manuscripts will be even more necessary in the future.

For now, if publishers allowed moderate use of chatbots, non-native English-speaking researchers could use them alongside traditional editorial services to refine the language.[47] The number of international researchers publishing in prestigious journals could, in the long run, indicate whether the scientific community has embraced the benefits of LLMs. However, we need to be cautious before drawing any conclusions from the sheer number of published papers. LLMs also create fertile ground for academic misconduct which increases the number of publications, like in the case of paper mills – multiple rewritings of the same paper.[48]

The question of how strict LLM policies should be is a matter of trust – whether we put our confidence in peers or the technology, we do not fully understand it, nor can we vouch for its reliability. The argument for putting more faith in our colleagues than AI tools for LLM detection is as epistemological as it is based on goodwill. From an ethical point of view, detection tools will be unproblematic only after we minimize the risk of false positives and ensure that they work equally well for native and non-native English speakers. From the perspective of epistemology, it is rational to give preference to our peers, as they – unlike AI applications – are not a black box. There are standards and procedures for testing claims and findings of other scholars. Ideally, we will manage to (re)establish epistemic trust in the scientific community[49] and approach our peers in the belief that they seek true answers, not instant gratification through reliance on unverified data. The path towards the fair use of LLMs in research, thus, requires broad discussions on responsibility, intellectual honesty, and the risks of relying on unverified data.

47 S.I. Hwang et al., *Is ChatGPT a "Fire of Prometheus" for Non-Native English-Speaking Researchers in Academic Writing?*, "Korean Journal of Radiology" 2023, Vol. 24, No. 10, 952, https://doi.org/10.3348/kjr.2023.0773.

48 G. Kendall, J.A. Teixeira da Silva, *Risks of Abuse of Large Language Models, Like ChatGPT, in Scientific Publishing: Authorship, Predatory Publishing, and Paper Mills*, "Learned Publishing" 2024, Vol. 37, No. 1, https://doi.org/10.1002/leap.1578.

49 W. Torsten, *Epistemic Trust in Science*, "British Journal for the Philosophy of Science" 2013, Vol. 64, No. 2, pp. 233–253, https://doi.org/10.1093/bjps/axs007.

## 6. Conclusions

LLM-based tools have changed the academic and scientific landscape. Laborious and time-consuming tasks, such as grammar checking and rare-literature searches, can now be assigned to machines, allowing researchers to focus more on intellectual pursuits. At the same time, the level of trust within the scientific community has decreased, as researchers may include AI-generated content in manuscripts. If unsanctioned, this trend could lead to numerous problems – from false authorship claims to unverified and incorrect data in scientific journals. In response, many academic institutions and publishers have banned LLMs to preserve the quality and integrity of research dissemination.

In this study we investigated whether such measures are justified and how their consequences unravel over time, especially for researchers who write in English but are not native speakers. We argue that the question of LLM restriction belongs in the discussion on linguistic privilege. AI detection tools not only report both false negatives and false positives, but non-native English speakers are more vulnerable to the latter due to their lower language proficiency. Labelling someone's paper as AI-generated warrants caution as it might harm their career and contribute to the linguistic privilege gap.

If academic institutions and scientific publishers continue to ban the use of LLMs, we risk forfeiting the benefits these technologies offer. LLM-based tools can help us mitigate linguistic disparity in the scientific community, as they offer learning opportunities, particularly for international researchers, who can use them for translation, paraphrasing, and grammar checking. However, even simple AI-generated essays require checking, as they may contain inaccuracies in terms of content and references. Additionally, these tools may not work as well in low-resource languages, and their reasoning skills are suboptimal. When LLMs improve in solving problems, a new challenge in verifying authorship will arise, as generated content will be even harder to detect.

From the epistemological point of view, the main concern is whether we can accurately distinguish AI-generated and human-written content. Relying on human judgement alone is insufficient, as we often fail to recognize whether LLMs were involved in manuscript writing. Studies that analyse the efficiency of AI tools for LLM detection reveal mixed results. Some of these tools are highly ac-

curate, but we encounter the black box problem. Both LLMs *and* AI tools we use to detect them need to become more transparent to earn our trust.

From an ethical perspective, the focus is on the impact of false positives, especially among international researchers. Relying on the discourse of linguistic epistemic injustice, we explored the concept of linguistic privilege. After that, we analysed some of the technologies in AI detection that contribute to a disproportionately higher rate of false positives among researchers who write in English as a second language.

Addressing the risks posed by LLMs is a task for the whole scientific community. The first step is to acknowledge the ethical and epistemic risk of putting too much trust in either LLMs or AI tools for their detection. We need more research on the differences in linguistic structures that native and non-native English speakers use. This could lead to further development of AI tools for LLM detection so they no longer target non-native speakers disproportionately. Employing these tools alongside human evaluation will help us avoid academic misconduct and maintain an inclusive approach. Finally, we should encourage a broad discussion on the long-term means of maintaining responsibility in science while enjoying the benefits of these technologies.

## Bibliography

Adetayo A.J., *Conversational Assistants in Academic Libraries: Enhancing Reference Services through Bing Chat*, "Library Hi Tech News" 2023, ahead of print, https://doi.org/10.1108/LHTN-08-2023-0142.

Aljamaan F., Temsah M.H, Altamimi I., Al-Eyadhy A., Jamal A., Alhasan K., Mesallam T.A., Farahat M., Malki K.H., *Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study*, "JMIR Medical Informatics" 2024, Vol. 12, No. 1, e54345, https://doi.org/10.2196/54345.

Alkaissi H., McFarlane S.I., *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*, "Cureus" 2023, Vol. 15, No. 2, e35179, pp. 1–4, https://doi.org/10.7759/cureus.35179.

Aydın Ö., Karaarslan E., *Is ChatGPT Leading Generative AI? What Is beyond Expectations?*, "Academic Platform Journal of Engineering and Smart Systems" 2023, Vol. 11, No. 3, pp. 118–134, https://doi.org/10.21541/apjess.1293702.

Berrami H., Jallal M., Serhier Z., Othmani M.B., *Exploring the Horizon: The Impact of AI Tools on Scientific Research*, "Data and Metadata" 2024, Vol. 3, https://doi.org/10.56294/dm2024289.

Bertin M., Atanassova I., Sugimoto C.R., Lariviere V., *The Linguistic Patterns and Rhetorical Structure of Citation Context: An Approach Using N-Grams*, "Scientometrics" 2016, Vol. 109, pp. 1417–1434, https://doi.org/10.1007/s11192-016-2134-8.

Cheng S.L., Tsai S.J., Bai Y.M., Ko C.H., Hsu C.W., Yang F.C., Tsai C.K., Tu Y.K., Yang S.N., Tseng P.T., Hsu T.W., *Comparisons of Quality, Correctness, and Similarity between ChatGPT-Generated and Human-Written Abstracts for Basic Research: Cross-Sectional Study*, "Journal of Medical Internet Research" 2023, Vol. 25, e51229, https://doi.org/10.2196/51229.

Day T., *A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT*, "The Professional Geographer" 2023, Vol. 75, No. 6, pp. 1024–1027, https://doi.org/10.1080/00330124.2023.2190373.

Desaire H., Chua A.E., Isom M., Jarosova R., Hua D., *Distinguishing Academic Science Writing from Humans or ChatGPT with Over 99% Accuracy Using Off-the-Shelf Machine Learning Tools*, "Cell Reports Physical Science" 2023, Vol. 4, No. 6, pp. 1–11, https://doi.org/10.1016/j.xcrp.2023.101426.

Dwivedi Y.K., Kshetri N., Hughes L., Slade E.L., Jeyaraj A., Kar A.K., Baabdullah A.M., Koohang A., Raghavan V., Ahuja M., Albanna H., *"So What if ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy*, "International Journal of Information Management" 2023, Vol. 71, 102642, https://doi.org/10.1016/j.ijinfomgt.2023.102642.

Elkhatat A.M., Elsaid K., Almeer S., *Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text*, "International Journal for Educational Integrity" 2023, Vol. 19, 17, https://doi.org/10.1007/s40979-023-00140-5.

Estévez-Ayres I., Callejo P., Hombrados-Herrera M.A., Alario-Hoyos C., Delgado Kloos C., *Evaluation of LLM Tools for Feedback Generation in a Course on Concurrent Programming*, "International Journal of Artificial Intelligence in Education" 2024, Vol. 35, pp. 774–790, https://doi.org/10.1007/s40593-024-00406-0.

Feuerriegel S., Hartmann J., Janiesch C., Zschech P., *Generative AI*, "Business & Information Systems Engineering" 2024, Vol. 66, No. 1, pp. 111–126, https://doi.org/10.1007/s12599-023-00834-7.

Fleckenstein J., Meyer J., Jansen T., Keller S.D., Köller O., Möller J., *Do Teachers Spot AI? Evaluating the Detectability of AI-Generated Texts among Student Essays*, "Computers and Education: Artificial Intelligence" 2024, Vol. 6, 100209, https://doi.org/10.1016/j.caeai.2024.100209.

Fricker M., *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, Oxford 2007.

Grindrod J., *Large Language Models and Linguistic Intentionality*, "Synthese" 2024, Vol. 204, 71, https://doi.org/10.1007/s11229-024-04723-8.

Guo D., et al., *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*, arXiv:2501.12948, https://doi.org/10.48550/arXiv.2501.12948.

Hannigan T.R., McCarthy I.P., Spicer A., *Beware of Botshit: How to Manage the Epistemic Risks of Generative Chatbots*, "Business Horizons" 2024, Vol. 67, No. 5, pp. 471–486, https://doi.org/10.1016/j.bushor.2024.03.001.

Hao Z., *Deep Learning Review and Discussion of Its Future Development*, "MATEC Web of Conferences" 2019, Vol. 277, 02035, https://doi.org/10.1051/matecconf/201927702035.

Harvard University, *Generative AI Guidance*, URL: https://oue.fas.harvard.edu/faculty-resources/generative-ai-guidance/.

Hiemstra D., *Language Models*, in: *Encyclopedia of Database Systems*, eds. L. Liu, M.T. Özsu, Springer, New York 2018, pp. 2061–2065, https://doi.org/10.1007/978-1-4614-8265-9_923.

Huang L., et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*, "ACM Transactions on Information Systems" 2024, Vol. 43, No. 2, 42, https://doi.org/10.1145/3703155.

Hwang S.I., Lim J.S., Lee R.W., Matsui Y., Iguchi T., Hiraki T., Ahn H., *Is ChatGPT a "Fire of Prometheus" for Non-Native English-Speaking Researchers in Academic Writing?*, "Korean Journal of Radiology" 2023, Vol. 24, No. 10, 952, https://doi.org/10.3348/kjr.2023.0773.

Jakesch M., Hancock J.T., Naaman M., *Human Heuristics for AI-Generated Language Are Flawed*, "Proceedings of the National Academy of Sciences" 2023, Vol. 120, No. 11, e2208839120, https://doi.org/10.1073/pnas.2208839120.

Kendall G., Teixeira da Silva J.A., *Risks of Abuse of Large Language Models, Like ChatGPT, in Scientific Publishing: Authorship, Predatory Publishing, and Paper Mills*, "Learned Publishing" 2024, Vol. 37, No. 1, https://doi.org/10.1002/leap.1578.

Lane H., Dyshel M., *Natural Language Processing in Action*, Manning Publications, Shelter Island 2025.

Liang W., Yuksekgonul M., Mao Y., Wu E., Zou J., *GPT Detectors Are Biased against Non-Native English Writers*, "Patterns" 2023, Vol. 4., No. 7, https://doi.org/10.1016/j.patter.2023.100779.

Ljubisavljevic D., Koprivica M., Kostic A., Devedžic V., *Homogeneity of Token Probability Distributions in ChatGPT and Human Texts*, "International Association for Development of the Information Society" 2023, pp. 207–213.

Makhortykh M., Baghumyan A., Vziatysheva V., Sydorova M., Kuznetsova E., *LLMs as Information Warriors? Auditing How LLM-Powered Chatbots Tackle Disinformation about Russia's War in Ukraine*, arXiv:2409.10697, https://doi.org/10.48550/arXiv.2409.10697.

Malik M.A., Amjad A.I., *AI vs AI: How Effective Are Turnitin, ZeroGPT, GPTZero, and Writer AI in Detecting Text Generated by ChatGPT, Perplexity, and Gemini?*, "Journal of Applied Learning and Teaching" 2024, Vol. 8, No. 1, https://doi.org/10.37074/jalt.2025.8.1.9.

Melliti M., *Using Genre Analysis to Detect AI-Generated Academic Texts*, "Diálogos" 2024, Vol. 16, No. 29, pp. 9–27, https://doi.org/10.61604/dl.v16i29.377.

Mirzadeh I., Alizadeh K., Shahrokhi H., Tuzel O., Bengio S., Farajtabar M., *GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models*, arXiv:2410.05229, https://doi.org/10.48550/arXiv.2410.05229.

Molina I.V., Montalvo A., Ochoa B., Denny P., Porter L., *Leveraging LLM Tutoring Systems for Non-Native English Speakers in Introductory CS Courses*, arXiv:2411.02725, https://doi.org/10.48550/arXiv.2411.02725.

Naveed H., Khan A.U., Qiu S., Saqib M., Anwar S., Usman M., Akhtar N., Barnes N., Mian A., *A Comprehensive Overview of Large Language Models*, arXiv:2307.06435, https://doi.org/10.48550/arXiv.2307.06435.

Obaidoon S., Wei H., *ChatGPT, Bard, Bing Chat, and Claude Generate Feedback for Chinese as Foreign Language Writing: A Comparative Case Study*, "Future in Educational Research" 2024, Vol. 2, No. 3, pp. 184–204, https://doi.org/10.1002/fer3.39.

Ozsoy M.G., *Multilingual Prompts in LLM-Based Recommenders: Performance across Languages*, arXiv:2409.07604, https://doi.org/10.48550/arXiv.2409.07604.

Picazo-Sanchez P., Ortiz-Martin L., *Analysing the Impact of ChatGPT in Research*, "Applied Intelligence" 2024, Vol. 54, pp. 4172–4188, https://doi.org/10.1007/s10489-024-05298-0.

Price G., Sakellarios M.D., *The Effectiveness of Free Software for Detecting AI-Generated Writing*, "International Journal of Teaching, Learning and Education" 2023, Vol. 2, No. 6, pp. 31–38, https://doi.org/10.22161/ijtle.2.6.4.

Ray P.P., *ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope*, "Internet of Things and Cyber-Physical Systems" 2023, Vol. 3, pp. 121–154, https://doi.org/10.1016/j.iotcps.2023.04.003.

Roumeliotis K.I., Tselikas N.D., *ChatGPT and Open-AI Models: A Preliminary Review*, "Future Internet" 2023, Vol. 15, No. 6, 192, https://doi.org/10.3390/fi15060192.

Russo F., Schliesser E., Wagemans J., *Connecting Ethics and Epistemology of AI*, "AI & Society" 2023, Vol. 39, pp. 1585–1603, https://doi.org/10.1007/s00146-022-01617-6.

Saarna C., *Identifying Whether a Short Essay Was Written by a University Student or ChatGPT*, "International Journal of Technology in Education" 2024, Vol. 7, No. 3, pp. 611–633, https://doi.org/10.46328/ijte.773.

Tang W., *Unlocking Second Language Students' Potential: ChatGPT's Pivotal Role in English for Academic Purposes Writing Success*, in: *Proceedings of the 2023 7th International Seminar on Education, Management and Social Sciences (ISEMSS 2023)*, Atlantis Press, 2023, pp. 694–706, https://doi.org/10.2991/978-2-38476-126-5_79.

Torsten W., *Epistemic Trust in Science*, "British Journal for the Philosophy of Science" 2013, Vol. 64, No. 2, pp. 233–253, https://doi.org/10.1093/bjps/axs007.

Tremblay A., Tucker B.V., *The Effects of N-Gram Probabilistic Measures on the Recognition and Production of Four-Word Sequences*, "The Mental Lexicon" 2011, Vol. 6, No. 2, pp. 302–324, https://doi.org/10.1075/ml.6.2.04tre.

Vishwanath P.R., et al., *Faithfulness Hallucination Detection in Healthcare AI*, in: *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.

Vučković A., Sikimić V., *Global Justice and the Use of AI in Education: Ethical and Epistemic Aspects*, "AI & Society", Vol. 40, pp. 3087–3104, https://doi.org/10.1007/s00146-024-02076-x.

Vučković A., Sikimić V., *How to Fight Linguistic Injustice in Science: Equity Measures and Mitigating Agents*, "Social Epistemology" 2022, Vol. 37, No. 1, pp. 80–96, https://doi.org/10.1080/02691728.2022.2109531.

Walters W.H., *The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors*, "Open Information Science" 2023, Vol. 7, No. 1, 20220158, https://doi.org/10.1515/opis-2022-0158.

Weber-Wulff D., Anohina-Naumeca A., Bjelobaba S., Foltýnek T., Guerrero-Dib J., Popoola O., Šigut P., Waddington L., *Testing of Detection Tools for AI-Generated Text*, "International Journal for Educational Integrity" 2023, Vol. 19, No. 1, pp. 26–65, https://doi.org/10.1007/s40979-023-00146-z.

Wolkovich E.M., *Obviously ChatGPT: How Reviewers Accused Me of Scientific Fraud*, "Nature," 5.02.2024, https://doi.org/10.1038/d41586-024-00349-5.

Wood P., *Oxford and Cambridge Ban ChatGPT over Plagiarism Fears but Other Universities Embrace AI Bot*, "The iPaper," 23.02.2023, URL: https://inews.co.uk/news/oxford-cambridge-ban-chatgpt-plagiarism-universities-2178391.

Ye H., Liu T., Zhang A., Hua W., Jia W., *Cognitive Mirage: A Review of Hallucinations in Large Language Models*, arXiv:2309.06794, https://doi.org/10.48550/arXiv.2309.06794.

Edukacja Filozoficzna
ISSN 0860-3839, eISSN 2956-8269
DOI: 10.14394/edufil.2025.0007
ORCID: 0000-0003-0115-5787

# Ethical Evaluation of Artificial Intelligence from the Perspective of the Catholic Church

Krzysztof Trębski
(Trnava University in Trnava, Slovakia)

**Abstract:** Artificial intelligence (AI) has emerged as a transformative force, profoundly reshaping many dimensions of human life. Its rapid growth, however, requires critical reflection on both benefits and risks. Ethical evaluation is not secondary but an opportunity to reconsider the meaning of human existence in a technology-driven world, while orienting progress with wisdom and foresight. The initial absence of clear frameworks has intensified debate on the urgent need for governance, legal safeguards, and moral principles to guide its invention, production, and use. This article analyzes the Catholic ethical evaluation of AI and the risks of unregulated development through documents of the Holy See, the teaching of recent popes, and their public pronouncements. It compares Catholic positions with existing governance instruments – such as the EU AI Act, UNESCO's *Recommendation on the Ethics of Artificial Intelligence*, and the *Rome Call for AI Ethics* with its *Hiroshima Addendum* – highlighting convergences and divergences, with particular attention to emerging ethical challenges. Based on the view that research and innovation are never morally neutral but always value-laden, the article underscores convergence between secular governance and Catholic teaching regarding the design, implementation, and responsible use of AI. At the same time, it highlights the Catholic emphasis on the centrality of the person – affirming that AI must serve humanity rather than replace or dominate it – on the inviolability of life (rejecting autonomous weapon systems), on human dignity (including principles such as non-discrimination, transparency, inclusion, accountability, reliability, safety, and privacy), on the dignity of work, social justice, and the universal call to fraternity. From this perspective, the Church supports a global ethical and regulatory framework, which it sees as essential not only to prevent harmful applications but also to promote virtuous practices and ensure continuous human oversight in the development and deployment of AI.

**Key words:** artificial intelligence, AI governance, ethical evaluation of AI in the Catholic Church

## 1. Introduction

We are witnessing the growing diffusion of artificial intelligence (AI), which elicits, on the one hand, uncritical enthusiasm and, on the other, excessive pes-

simism towards a tool that is at once "an exciting and fearsome tool,"[1] capable of generating immense benefits but also posing serious risks. This dual potential renders AI an inherently ambivalent system: it could become the most powerful multiplier of knowledge, bridging distances among people; yet it could equally evolve into a driver of injustice and social stratification. To prevent AI from becoming a multiplier of inequality – both between technologically advanced and developing nations, and between dominant and marginalized social groups – its development and implementation must be guided by robust political and ethical oversight.[2] Without such governance, AI risks undermining the "culture of solidarity and encounter," which is grounded in inclusion and dialogue,[3] and instead promoting a "culture of waste"[4] that fosters discrimination and marginalization.

## 2. Artificial Intelligence between Techno- and Human-Centrism

Technology, and particularly AI, with its capacity to shape material reality, mitigate risks, ease human labour, and enhance living conditions, embodies the objective dimension of human action. It must, however, be remembered that technology is not merely a human activity; rather, human nature itself constitutes a techno-human condition, insofar as the technical dimension is an intrinsic aspect of being human, an expression of existence as an individual, relational, and transcendent being.[5]

[1] The expression "an exciting and fearsome tool" was used by Pope Francis to emphasize that it is precisely the powerful technological progress that makes AI both a fascinating and a fearsome tool, calling for a level of reflection capable of meeting the challenge it presents. Cf. Francis, *Address of His Holiness Pope Francis*, Borgo Egnazia, 14.06.2024, URL: https://www.vatican.va/content/francesco/en/speeches/2024/june/documents/20240614-g7-intelligenza-artificiale.html.

[2] S. Quintarelli et al., *AI: profili etici. Una prospettiva etica sull'Intelligenza Artificiale. Princìpi, diritti e raccomandazioni*, "BioLaw Journal – Rivista di BioDiritto" 2019, Vol. 3, pp. 183–204.

[3] Cf. Francis, *Message of Pope Francis for the 48th World Communications Day: Communication at the Service of an Authentic Culture of Encounter*, 1.06.2014, URL: https://www.vatican.va/content/francesco/en/messages/communications/documents/papa-francesco_20140124_messaggio-comunicazioni-sociali.html; Francis, *Address of Holy Father Francis*, Cagliari, 22.09.2013, URL: https://www.vatican.va/content/francesco/en/speeches/2013/september/documents/papa-francesco_20130922_cultura-cagliari.html.

[4] Cf. Francis, *General Audience*, Saint Peter's Square, 5.06.2013, URL: https://www.vatican.va/content/francesco/en/audiences/2013/documents/papa-francesco_20130605_udienza-generale.html.

[5] Cf. P. Benanti, *Homo Faber: The Techno-Human Condition*, EDB, 2018, pp. 108, 110, 112.

This integral anthropological vision underscores the need for ongoing discernment to ensure that AI does not reduce the human being to a mere instrument of efficiency or productivity, but rather recognizes and safeguards the inalienable dignity of every person. Technology is born with a purpose and, through its interaction with human society, always represents a form of ordering social relations and a structure of power – empowering some to act while restricting others. This constitutive dimension of power inherently carries, whether explicitly or implicitly, the worldview of its creators and developers.[6]

Proponents of a techno-centric vision of development, who advocate for every form of technologization of the body and mind, envisage horizons in which the artificial becomes increasingly indistinguishable from the natural, intentionally erasing the difference between human and machine in a symbiotic fusion of humanity and technology, of organic and inorganic life. They promote the advancement of convergent technologies and robotics/AI, wherein the robot serves as the embodiment of AI, designed to replace and ultimately surpass the human being.[7] This is presented as the sole path towards overcoming the biological limitations of the body and the neurocognitive constraints of the mind, thereby moving towards a trans-human, post-human, or even super-human perfection. If the techno-centric worldview were to prevail, good would ultimately be reduced to what can be technologically achieved. In such a framework – where efficiency and utility become the sole criteria of judgement – authentic development is inevitably denied. True development, in fact, cannot be reduced merely to "doing." Its key lies in a mind capable of grasping the fully human meaning of action within a holistic vision of being.[8] Even when AI is employed, fundamental decisions remain human in nature and therefore require moral responsibility. There are strong anthropological, ontological, and ethical reasons to affirm that the non-reproducibility, non-substitutability, and uniqueness of human intelligence constitute a higher value.[9]

---

6 Cf. L. Winner, *Do Artifacts Have Politics?*, in: L. Winner, *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, Chicago 1988, p. 23.

7 Cf. E. Sadin, *Critica della ragione artificiale. Una difesa dell'umanità*, Luiss University Press, Milano 2019, pp. 10–33.

8 Cf. Francis, *Address Prepared by Pope Francis, Read by H.E. Archbishop Paglia, President of the Pontifical Academy for Life, Meeting with the Participants in the Plenary Assembly of the Pontifical Academy for Life*, Vatican City, 28.02.2020, https://www.vatican.va/content/francesco/en/speeches/2020/february/documents/papa-francesco_20200228_accademia-perlavita.html.

9 See L. Floridi, J.W. Sanders, *Artificial Evil and the Foundation of Computer Ethics*, "Ethics and Information Technology" 2001, Vol. 3, No. 1, pp. 55–66.

In the current scientific context, marked by the expanding presence of AI in vast domains of human activity, it becomes indispensable to develop a critical philosophical reflection on the human being – its meaning and value – in order to identify the potential limits of technology.[10] The challenge is not to exalt technology while disparaging the human person, nor to exalt the human while rejecting technology. Rather, the objective is to enable interventions upon the human condition without distorting its identity and without triggering irreversible transformations. In this sense, the task is not merely to acknowledge what remains human despite technology, but above all to discern what must remain human through technology.[11] If we understand the limits of what we can do with technology, we can make better choices about what we should do with it to make the world better for everyone.[12]

## 3. Core Ethical Principles in the Age of Artificial Intelligence

Given the vast scope of the phenomenon of AI and the significant progress achieved by such systems, many have sought to propose various initiatives aimed at defining the principles that should underlie AI, which must be viewed from a perspective that benefits humanity.

The four foundational principles of biomedical ethics – autonomy, beneficence, non-maleficence, and justice – developed by Tom L. Beauchamp and James F. Childress and first introduced in 1979,[13] embody fundamental moral values shared by individuals committed to ethical conduct and can therefore be regarded as central pillars in discussions on the ethical foundations that should guide the design, development, and use of AI.[14]

[10] Cf. T. Hagendorff, *The Ethics of AI Ethics: An Evaluation of Guidelines*, "Minds & Machines" 2020, Vol. 30, pp. 99–120.

[11] Cf. L. Floridi, J.W. Sanders, *On the Morality of Artificial Agents*, "Minds & Machines" 2004, Vol. 14, No. 3, pp. 349–379.

[12] Cf. M. Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, The MIT Press, Cambridge, MA–London 2019, p. 12.

[13] T.L. Beauchamp, J.F. Childress, *Principles of Biomedical Ethics*, Oxford University Press, Oxford 1979.

[14] Beauchamp and Childress maintain that these norms have developed because the essential role of morality as a social institution is to support human flourishing by addressing the factors that diminish well-being and by preventing conditions such as indifference, conflict, suffering, hostility, scarcity, and misinformation. Historical evidence demonstrates that when such moral

The principle of autonomy recognizes the capacity of individuals to self-determine and to act according to their own moral values and convictions. It implies that every person must be able to exercise meaningful control over their choices, remaining free from external coercion and internal constraints that could compromise voluntariness and understanding.[15] As Beauchamp and Childress explain, autonomy is self-rule that is free from both controlling interference by others and from limitations, such as inadequate understanding, that prevent meaningful choice.[16] In this sense, the principle is expressed in the power to decide, including the power to choose whether and when to decide.[17] Such a capacity constitutes the core of moral self-determination and forms the foundation of all respect for human dignity.[18] In the context of AI ethics, the principle of autonomy acquires growing significance, as intelligent systems increasingly interact with human decision-making processes. Ethically sound AI design must therefore aim to preserve – and, where possible, enhance – human capacities for comprehension, deliberation, and informed decision-making. This entails ensuring that users understand how AI systems operate, what data they use, and how their outputs are generated, so that individuals can make genuinely voluntary and informed choices regarding their interaction with these systems.[19] Ultimately, respecting autonomy in the age of AI means promoting a balanced relationship between humans and machines – one in which AI serves as a tool for cognitive and decision-making empowerment, rather than as a replacement for human will or moral responsibility.

The principle of beneficence ("do good only")[20] mandates that AI be developed and applied with the primary objective of generating tangible benefits for individu-

norms are ignored, human life deteriorates into misery, violence, and distrust. Conversely, respecting and upholding these norms helps to reduce suffering and promote social harmony. Therefore, they are vital for improving human well-being and achieving the fundamental aims of morality. Cf. T.L. Beauchamp, *Standing on Principles: Collected Essays*, Oxford University Press, New York 2010, pp. 43–44.

15 Cf. T.L. Beauchamp, J.F. Childress, *Principles of Biomedical Ethics*, 8th ed., Oxford University Press, New York–Oxford 2019, pp. 99–111.

16 Cf. ibid., p. 101.

17 Cf. S. Hajkowicz, *Global Megatrends: Seven Patterns of Change Shaping Our Future*, CSIRO Publishing, Melbourne 2015, p. 91.

18 Cf. P. Lin, K. Abney, G. Bekey, *Robot Ethics: Mapping the Issues for a Mechanized World*, "Artificial Intelligence" 2011, Vol. 175, Nos. 5–6, pp. 942–949, https://doi.org/10.1016/j.artint.2010.11.026.

19 Cf. L. Floridi et al., *AI 4 People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, "Minds & Machines" 2018, Vol. 28, No. 4, p. 698.

20 Cf. T.L. Beauchamp, J.F. Childress, *Principles of Biomedical Ethics*, 8th ed., op. cit., p. 217.

als and society as a whole.[21] This principle encompasses three fundamental dimensions: the promotion of well-being, the safeguarding of the intrinsic dignity of every person, and the sustainability of technological development, which includes the protection of the environment.[22] Specifically, the promotion of well-being entails that AI should contribute meaningfully to improving the quality of human life by enhancing cognitive, relational, and operational capacities, while simultaneously reducing social and economic inequalities. The protection of human dignity constitutes a second essential dimension of beneficence: every application of AI must respect and value the human being as an end in itself, avoiding any form of objectification, manipulation, or algorithmic discrimination. Ethically oriented AI must therefore be conceived as a tool of human empowerment – one that supports decision-making and action without replacing individual will or moral responsibility. Finally, beneficence requires a sustained commitment to sustainability, understood as a balance between technological progress and environmental responsibility. The development and deployment of AI systems should be designed to ensure efficient resource use, minimize ecological impact, and promote an innovation model that does not compromise the well-being of future generations.

Overall, the principle of beneficence, when applied to the domain of AI, calls for an ethical vision oriented towards the "digital common good," in which technology functions as an enabling force for the promotion of human welfare, the protection of dignity, and the preservation of the environment. Only within this framework can AI be regarded not merely as technologically advanced, but also as morally justifiable and socially sustainable.[23]

The principle of non-maleficence ("do no harm") requires the deliberate avoidance of actions that may cause harm to individuals or society.[24] It thus establishes

[21] Cf. A. Jobin, M. Ienca, E. Vayena, *The Global Landscape of AI Ethics Guidelines*, "Nature Machine Intelligence" 2019, Vol. 9, No. 1, pp. 389–399.

[22] Cf. M. Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Data & Society, URL: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

[23] Cf. L. Floridi et al., *AI 4 People*, op. cit. In particular, see point 4.1: "Beneficence: promoting well-being, preserving dignity, and sustaining the planet."

[24] Lorenzo D'Avack combined the principle of beneficence with that of non-maleficence, explaining that such systems, in addition to contributing to the improvement of human well-being, should also avoid causing harm to individuals and society. Cf. L. D'Avack, *La rivoluzione tecnologica e la nuova era digitale: problemi etici*, in: *Intelligenza Artificiale. Il diritto, i diritti, l'etica*, ed. U. Ruffolo, Giuffrè, Milano 2020, pp. 3–28.

a minimal threshold of ethically acceptable behaviour, below which practices become detrimental to the dignity or integrity of the person. As Beauchamp and Childress emphasize,[25] non-maleficence highlights the moral obligation not only to refrain from intentionally causing harm but also to anticipate and prevent potential risks that may arise from technological or procedural decisions. In the context of AI, this principle assumes particular significance in three key domains: privacy, security, and capability caution. First, with regard to privacy, AI systems must not violate the right to personal data protection or intrude upon individuals' private spheres, as such violations would constitute a direct harm to autonomy and human dignity.[26] Second, concerning security, AI systems must be designed to ensure robustness, reliability, and resistance to malicious use, errors, or unintended consequences that could result in physical, psychological, or social harm. Preventing malfunctions and ensuring safety therefore represent essential components of ethically responsible AI design. Finally, the concept of capability caution refers to the responsibility of avoiding the development or deployment of systems whose capacities could become dangerous if they were to exceed or escape human control. This includes both the containment of potentially harmful autonomous functions and the governance of AI systems whose operational scope may produce unforeseen or uncontrollable effects.[27]

The principle of justice concerns the promotion of prosperity and the preservation of solidarity within society. AI must function as an instrument to reduce, not exacerbate, social and economic inequalities, ensuring that its benefits are distributed fairly and that no one is left behind.[28] In this sense, justice in the domain of AI – understood as impartiality – is best described through the concept of "algorithmic fairness."[29]

According to Luciano Floridi and Josh Cowls,[30] the framework of the four principles derived from bioethics should be supplemented with a fifth principle,

---

25 Cf. T.L. Beauchamp, J.F. Childress, *Principles of Biomedical Ethics*, 8th ed., op. cit., pp. 133–136.

26 Cf. L. Floridi et al., *AI 4 People*, op. cit. In particular, see point 4.2: "Non-maleficence: privacy, security and 'capability caution.'"

27 Cf. A. Jobin, M. Ienca, E. Vayena, *The Global Landscape of AI Ethics Guidelines*, op. cit., p. 392.

28 Cf. L. Floridi et al., *AI 4 People*, op. cit. In particular, see point 4.4: "Justice: promoting prosperity and preserving solidarity."

29 Cf. J. Morley et al., *Ethics as a Service: A Pragmatic Operationalisation of AI Ethics*, "Minds & Machines" 2021, Vol. 31, No. 2, pp. 239–356, https://doi.org/10.1007/s11023-021-09563-w.

30 Cf. L. Floridi, J. Cowls, *Unified Framework of Five Principles for AI in Society*, "Harvard Data Science Review" 2019, Vol. 1, pp. 2–15.

explicability, specifically designed to address the unique ethical challenges posed by AI systems. This principle is crucial because it enables the effective implementation of all other ethical principles.[31] Given that AI systems are often characterized by significant technical and conceptual opacity, explicability encompasses two complementary dimensions: intelligibility, that is, the capacity to understand how a system functions ("How does it work?"), and accountability, understood as the ability to identify who is responsible for the system's functioning and its consequences ("Who is responsible for the way it works?").[32] There is broad consensus that accountability with respect to moral and legal norms, as well as the associated liability, represents an essential requirement for any AI technology. The central issue, however, particularly concerning autonomous systems and robots with independent decision-making capacities, is how such responsibility can be effectively ensured and how moral and legal accountability can be assigned in the event of unintended or harmful outcomes.[33]

In this context, the principle of explicability goes beyond promoting technical transparency; it constitutes a necessary condition for ensuring public trust, the traceability of algorithmic decisions, and the ethical and legal legitimacy of AI deployment in contemporary society.

Building on these principles, it becomes necessary to define how AI research should develop so as not to harm humanity: it must remain under human control, be designed transparently and intelligibly, and be developed and applied fairly, in such a way that it neither perpetuates nor exacerbates existing inequalities.[34] A central challenge lies in the difficulty of achieving full transparency in the decision-making processes of AI systems based on deep neural networks. For this reason, a balance must be sought between the efficiency of results and their interpretability. Through the systematic recording and ongoing analysis of AI actions, it is possible to verify their compliance with ethical and legal principles, to

[31] Cf. L. Floridi et al., *AI 4 People*, op. cit. In particular, see point 4.5: "Explicability: Enabling the other principles through intelligibility and accountability."

[32] Cf. J.A. Kroll et al., *Accountable Algorithms*, "University of Pennsylvania Law Review" 2017, Vol. 165, No. 3, p. 645.

[33] Cf. J. Morley et al., *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" 2020, Vol. 26, No. 4, pp. 2141–2168, https://doi.org/10.1007/s11948-019-00165-5.

[34] Cf. M. Taddeo, L. Floridi, *How AI Can Be a Force for Good: An Ethical Framework Will Help to Harness the Potential of AI while Keeping Humans in Control*, "Science" 2018, Vol. 361, No. 6404, pp. 751–752.

identify and correct potential biases or errors, and to strengthen user trust. This process not only improves AI models but also ensures that their development remains ethical and sustainable.[35] To this end, it is crucial to distinguish between AI "decisions," which can be traced back to computational activity, and human "choices."[36] The latter require profound ethical reflection, drawing upon history, culture, and a shared system of values, since every act of choosing is the product of judgement rather than mere calculation.[37] It is therefore indispensable that human beings establish the boundaries and rules necessary to guarantee a responsible use of this technology – one that should always serve the highest potential and aspirations of humankind,[38] while safeguarding those human functions that cannot and must not be replaced by machines: judgement, respect, understanding, caring, and love.[39]

## 4. Secular Models of Artificial Intelligence Governance

The accelerated evolution of AI technologies has given rise to profound ethical, social, and legal challenges, thereby necessitating the establishment of robust and coherent governance frameworks.[40] In this context, instruments such as UNESCO's

[35] Cf. L. Floridi, *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford University Press, Oxford 2023, pp. 105–112.

[36] Cf. L. Floridi, F. Cabitza, *Intelligenza artificiale. L'uso delle nuove macchine*, Bompiani, Firenze–Milano 2021, p. 70.

[37] Cf. D.M. Berry, *The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot*, "Weizenbaum Journal of the Digital Society" 2023, Vol. 3, No. 3, https://doi.org/10.34669/WI.WJDS/3.3.2.

[38] Francis, *Message of His Holiness Pope Francis for the 57th World Day of Peace: Artificial Intelligence and Peace*, 1.01.2024, URL: https://www.vatican.va/content/francesco/it/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html.

[39] Cf. J. Weizenbaum, *Il potere del computer e la ragione umana. I limiti dell'intelligenza artificiale*, EGA-Edizioni Gruppo Abele, Torino 1987, p. 192.

[40] Floridi clarifies the term governance and emphasizes that digital governance, digital ethics (also known as computer, information, or data ethics), and digital regulation represent distinct normative approaches. Digital governance refers to the practice of defining and implementing policies, procedures, and standards for the proper development, use, and management of the infosphere. It may include guidelines and recommendations that overlap with digital regulation, without necessarily coinciding entirely with it. Digital regulation, on the other hand, refers to the system of laws developed and enforced by social or governmental institutions to regulate the behaviour of agents. Not every aspect of digital regulation pertains to digital governance, and not every aspect of digital governance falls under regulation. Floridi highlights the need for

*Recommendation on the Ethics of Artificial Intelligence*[41] and the European Union's AI Act[42] represent two pivotal regulatory models. While differing in scope and legal enforceability, both initiatives converge on a set of foundational ethical principles, thereby contributing to the broader debate on global AI governance. Their significance lies not only in establishing normative ethics standards for the responsible development and deployment of AI[43] but also in fostering international dialogue aimed at reconciling diverse ethical traditions and regulatory approaches in the pursuit of a shared, human-centred digital future.[44]

As a transformative force, AI gives rise to global ethical, social, and political questions. In this context, UNESCO's *Recommendation on the Ethics of Artificial Intelligence*, adopted unanimously in November 2021 by all 193 Member States, represents a significant attempt to establish a shared international framework. It identifies four foundational values: human dignity and human rights, social justice, inclusiveness, and environmental sustainability.[45] These values underpin the formulation of guiding principles and policy actions intended to ensure that AI development and deployment serve the common good while respecting fundamental rights. Human dignity occupies a central place in the *Recommendation*, understood as the intrinsic and equal worth of every individual, which cannot be compromised at any stage of the AI lifecycle. Technologies must therefore contribute to enhancing human well-being without objectifying, subordinating, or discriminating against individuals or communities, with particular attention to vulnerable groups.[46] Environmental protection constitutes another key principle,

ethical guidance in the governance of AI. Cf. L. Floridi, *Soft Ethics, the Governance of the Digital and the General Data Protection Regulation*, "Philosophical Transactions of the Royal Society A" 2018, Vol. 376, No. 2133, https://doi.org/10.1098/rsta.2018.0081.

41 UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris 2022, URL: https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000381137&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_75c9fb6b-92a6-4982-b772-79f540c9fc39%3F_%3D381137eng.pdf&locale=en&multi=true&ark=/ark:/48223/pf0000381137/PDF/381137eng.pdf#1517_21_EN_SHS_int.indd%3A.8946%3A.

42 European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council*, 13.07.2024, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

43 V.C. Müller, *Ethics of Artificial Intelligence and Robotics*, in: *Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), ed. E.N. Zalta, URL: https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/.

44 M. Coeckelbergh, *AI Ethics*, The MIT Press, Cambridge, MA, 2020, p. 57.

45 UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, op. cit. p. 8.

46 Cf. ibid., p. 10.

as ecosystems are considered indispensable for human welfare and for future generations. Stakeholders involved in AI development and use are required to minimize environmental impacts through sustainable practices and adherence to the precautionary principle.[47] The *Recommendation* further underscores the importance of inclusion and diversity, which must be safeguarded by avoiding social, digital, or cultural exclusion and by promoting the active participation of all groups, regardless of origin, gender, age, religion, disability, or socio-economic condition.[48] It also stresses the need to foster peaceful, just, and interconnected societies in which AI serves as a tool for solidarity, justice, and equity, without undermining human autonomy or fuelling social or environmental conflicts.[49] Among its operational principles, the document highlights proportionality and the imperative to "do no harm," restricting AI applications to legitimate and proportionate purposes, particularly in contexts directly affecting human life and death.[50] It also requires safety and security mechanisms to prevent risks and vulnerabilities, fair access to the benefits of AI, and continuous assessment of the social, economic, and environmental consequences of technology. Further principles include the protection of privacy and personal data through adequate regulatory frameworks, human oversight and accountability – ensuring that ultimate responsibility rests with natural or legal persons – together with transparency and explainability as essential conditions for trust, traceability, and avenues of redress.[51] The *Recommendation* emphasizes the importance of digital literacy and public awareness, enabling citizens and communities to understand the implications of AI and make informed choices. It calls for a multilevel, collaborative, and adaptive governance model engaging governments, civil society, the private sector, academia, and local communities, in full respect of cultural diversity and territorial specificities.[52] In addition, clear requirements for transparency and explainability must be complemented by measures to counteract bias and stereotypes in datasets. Diversity and inclusion in technological development and access should be actively promoted, while States are encouraged to contribute to the formulation of international standards ensuring safety, reliability, and respect for

47 Cf. ibid., p. 12.
48 Cf. ibid., p. 16.
49 Cf. ibid., pp. 22–25.
50 Cf. ibid., p. 20.
51 Cf. ibid., p. 8.
52 Cf. ibid., p. 21.

human dignity. With regard to data governance, quality, security, and protection are paramount, together with corrective feedback mechanisms. Privacy safeguards should be rooted in privacy by design, impact assessments, and legislation aligned with international law, ensuring that individuals retain full control over their personal data, including rights of access, erasure, and enhanced protection for sensitive categories such as biometric, genetic, and health information.[53]

The main criticisms of UNESCO's *Recommendation on the Ethics of Artificial Intelligence* focus on both theoretical and practical limitations. First, scholars emphasize its non-binding character: although it represents the first global attempt to establish a shared ethical framework, it lacks legal force and delegates the responsibility for implementation to Member States. This feature raises doubts about its operational effectiveness, particularly in political contexts where AI governance does not constitute a strategic priority.[54] A second critical point concerns the generality of the principles, which are often formulated in broad and indeterminate terms. While this vagueness facilitates international consensus, it risks undermining the translation of these principles into concrete guidelines and regulatory mechanisms.[55] Moreover, the *Recommendation* fails to adequately address emerging issues, such as the legal responsibility of autonomous systems, the impact of generative technologies, and the geopolitical challenges linked to data sovereignty.[56] For these reasons, the document is often regarded as a preliminary ethical framework: valuable as a general point of reference, yet insufficient to govern the complexity of the ongoing transformations.

The European Regulation on Artificial Intelligence (EU AI Act)[57] constitutes the first comprehensive attempt to regulate AI systems within the European Union, establishing a normative framework designed to reconcile technological innovation with the protection of fundamental rights. It is inspired by the principles developed by the European Commission's High-Level Expert Group on AI,[58]

53 L. Floridi, *The Ethics of Artificial Intelligence*, op. cit., pp. 112–115.

54 L. Floridi, J. Cowls, *A Unified Framework of Five Principles for AI in Society*, in: L. Floridi, ed., *Ethics, Governance, and Policies in Artificial Intelligence*, Springer Verlag, Cham, 2021, p. 15.

55 A. Jobin, M. Ienca, E. Vayena, *The Global Landscape of AI Ethics Guidelines*, op. cit., p. 392.

56 C. Cath, *Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges*, "Philosophical Transactions of the Royal Society A" 2018, Volume 376, No. 2133, 20180080, https://doi.org/10.1098/rsta.2018.0080.

57 European Union, *Regulation (EU) 2024/1689*, op. cit.

58 High-Level Expert Group on AI (AI HLEG), *Ethics Guidelines for Trustworthy AI*, 8.04.2019, URL: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

which identified four fundamental ethical principles regarded as the foundation of trustworthy AI: (1) respect for human autonomy; (2) prevention of harm; (3) fairness; and (4) explicability. However, in order to effectively achieve reliable AI, they outlined seven key prerequisites that, in their view, must be continuously monitored and managed throughout the entire lifecycle of AI systems: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental well-being; and (7) accountability. Furthermore, the group emphasized the potential necessity of introducing new legal measures and control mechanisms capable of ensuring adequate protection against negative effects, while enabling effective human ethical oversight in the processes of design, development, and deployment of AI technologies.

Among its most significant aspects, the AI Act introduces a regime of explicit prohibitions targeting practices deemed incompatible with human dignity and collective security.[59] AI systems that may adversely affect safety or fundamental rights are classified as "high-risk" under the EU AI Act. This category encompasses, on the one hand, systems integrated into products already subject to EU product safety legislation, such as toys, aviation technologies, motor vehicles, medical devices, and lifts. On the other hand, it includes applications operating in sensitive domains, such as critical infrastructure management, education and vocational training, employment and labour relations, access to essential private and public services, law enforcement, migration and border control, as well as systems used in legal interpretation and application.[60] Similarly, the regulation bans the use of technologies exploiting vulnerabilities related to age, disability, or socio-economic conditions, where such exploitation results in behavioural distortion with damaging consequences.

A further prohibition concerns social scoring mechanisms, namely the classification of individuals based on behaviours or personal characteristics. This practice is considered harmful, as it may generate discriminatory or dispropor-

[59] For further developments on the topic, see the 2025 updates: European Commission, *Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act)*, C(2025) 5052 final, 29.07.2025, URL: https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.

[60] European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, COM(2021)206final, 21.04.2021, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

tionate treatment, particularly when applied in contexts different from those in which the data were originally collected.[61] Likewise, predictive systems that assign criminal risk to individuals solely on the basis of automated profiling are banned, with the exception of tools that support human evaluation grounded in objective and verifiable evidence.

The regulation also restricts the creation of biometric databases through indiscriminate scraping of facial images from the internet or surveillance systems, a practice deemed invasive of privacy and likely to foster mass surveillance. Similarly, it prohibits the use of systems intended to infer emotional states in professional or educational contexts, except in narrowly defined medical or security circumstances. Furthermore, biometric categorization aimed at deducing sensitive attributes – such as race, religious belief, sexual orientation, or political opinion – is forbidden, with exceptions limited to legitimate purposes, like dataset labelling for research or security activities.

Equally significant are the obligations imposed on generative and general-purpose models, which must provide adequate technical documentation, comply with copyright law, and disclose transparency regarding training data.[62] These provisions are designed to mitigate risks associated with violations of fundamental rights while reinforcing public trust through enhanced traceability of decision-making processes. Ultimately, the EU AI Act represents an innovative regulatory model capable of translating ethical principles into binding legal obligations, thereby consolidating a European approach centred on human dignity, fairness, and sustainability. It serves as a bridge between ethical reflection and political action, fostering a digital ecosystem where innovation is guided by the common good and respect for fundamental values.

## 5. Catholic Church's Vision of Artificial Intelligence

The development of AI in contemporary society represents one of the most profound ethical and anthropological challenges of our time. In this context, there

[61] A. Atabekov, A. Yastrebov, *Legal Status of Artificial Intelligence across Countries: Legislation on the Move*, "European Research Studies Journal" 2018, Vol. 21, No. 4, pp. 773–782.

[62] European Commission, *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, 19.02.2020, URL: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

emerges an urgent need to formulate an ethics of discernment and decision-making that, in the light of Catholic teaching, reaffirms the primacy of the spirit over matter[63] and ensures that technology remains at the service of the human person, rather than becoming its master. The defining risk of our age lies in the emergence of a "technocratic paradigm," a worldview that tends to subordinate the human person to the power of machines and the logic of efficiency, thereby obscuring the spiritual, moral, and relational dimensions of human existence.[64] It is therefore essential to understand these profound transformations and to orient them towards serving the human person, while safeguarding and promoting inherent human dignity. Given the complexity and unpredictability of such developments, this task calls for particularly deep ethical discernment.[65]

The Catholic Church's support for the ethical moderation of algorithms reflects an awareness that, given the complexity of today's technological landscape, a more sophisticated ethical framework is required to ensure that this commitment is genuinely effective.[66] It is therefore essential to maintain a robust ethical framework throughout the entire process of AI development – from design to deployment and use – in order to guide the values shaping this ongoing transformation for the common good.[67] From this necessity arises the proposal of

---

63 Cf. Benedict XVI, *Encyclical Letter Caritas in Veritate*, Rome, 29.06.2009, par. 69–70, URL: https://www.vatican.va/content/benedict-xvi/en/encyclicals/documents/hf_ben-xvi_enc_20090629_caritas-in-veritate.html.

64 Cf. Francis, *Apostolic Exhortation Laudate Deum*, Rome, 4.10.2023, par. 21, URL: https://www.vatican.va/content/francesco/en/apost_exhortations/documents/20231004-laudate-deum.html.

65 Cf. Francis, *Letter of His Holiness Pope Francis to the President of the Pontifical Academy for Life for the 25th Anniversary of the Establishment of the Academy: Humana Communitas*, Vatican City, 6.01.2019, par. 12, URL: https://www.vatican.va/content/francesco/en/letters/2019/documents/papa-francesco_20190106_lettera-accademia-vita.html.

66 Cf. Francis, *Address of His Holiness Pope Francis to the Participants in the Congress on "Child Dignity in the Digital World"*, Vatican City, 6.10.2017, URL: https://www.vatican.va/content/francesco/en/speeches/2017/october/documents/papa-francesco_20171006_congresso-child-dignity-digitalworld.html.

67 Cf. Pontifical Academy for Life, *Rome Call for AI Ethics*, Rome, 28.02.2020, URL: https://www.vatican.va/roman_curia/pontifical_academies/acdlife/documents/rc_pont-acd_life_doc_20202228_rome-call-for-ai-ethics_en.pdf. The *Rome Call for AI Ethics* is a document promoting a shared ethical approach to AI. It aims to ensure that digital innovation and technological progress serve humanity by putting the human person at the centre. The signatories advocate for a new "algor-ethics" to guide the development of AI that respects human dignity, benefits everyone, and does not focus solely on profit or the replacement of workers.

algor-ethics,[68] a fully human and responsible approach to AI, as promoted by the Catholic Church.

Algor-ethics, understood as applied ethics in the field of AI, requires an assessment not only of the ways in which AI models are designed, developed, and used by human beings, but also of the social and environmental impacts that these systems may exert on society and the natural environment through their operation and behaviour. It thus assumes a dual nature. On the one hand, it seeks to identify the principles that human beings must observe to ensure that AI systems are developed exclusively to promote sustainable social well-being, adopting not merely a technical approach but a multidisciplinary one that integrates perspectives from computer science, engineering, psychology, anthropology, philosophy, religion, and political science. On the other hand, algor-ethics also represents an attempt to encode within AI systems a set of behavioural rules that enable machines to act in ways that respect the human person.

In this context, the global initiative *Rome Call for AI Ethics*[69] – launched by the Pontifical Academy for Life (Holy See, Vatican) with the support of the RenAIssance Foundation,[70] established by Pope Francis on 12 April 2021 to promote an ethical approach to the development and use of artificial intelligence worldwide – assumes particular significance. The aim of this initiative was to propose, with broad international and interfaith consensus, that AI development should adopt, from the very beginning of algorithm design, an "algor-ethical" approach – ethics integrated into the design itself, or "ethics by design." This effort seeks to promote algor-ethics, ensuring that AI is used in an ethical manner. To this end, the *Rome Call for AI Ethics* proposes six ethical evaluation criteria for AI: transparency, inclusion, responsibility, impartiality, reliability, and respect for security and privacy, so that AI benefits all individuals and safeguards human dignity.[71] In light of the *Rome Call*, which articulates the Catholic Church's position

[68] See P. Benanti, *Oracoli. Tra algoretica e algocrazia*, Luca Sossella Editore, Roma 2018.

[69] Pontifical Academy for Life, *Rome Call for AI Ethics*, op. cit.

[70] Cf. RenAIssance Foundation, URL: https://www.romecall.org/renaissance-foundation/.

[71] The contribution of Floridi and Cowls influenced the six AI governance principles proposed in the *Rome Call*. The explainability principle proposed by Floridi clearly shaped the content of the document, as he was directly involved in its development. However, it is important to emphasize that the five original principles elaborated by Floridi and Cowls do not fully coincide with the six principles set forth in the *Rome Call*.

on AI,[72] shared ethical principles acquire crucial importance in addressing contemporary challenges.[73] Foremost among these is the need for transparency, ensuring that all machine-generated content is immediately recognizable. This principle is closely linked to accountability, which requires establishing standards to trace the origin and authenticity of digital content, thereby countering the spread of disinformation and fake news.

Moreover, the development of AI systems must prioritize inclusivity, respecting the diversity of cultures, traditions, and languages that define humanity. This entails a strong commitment to fairness, ensuring that generative AI does not perpetuate or amplify existing biases. Given their far-reaching societal impact, the reliability and robustness of such systems are of primary importance.[74] Finally, safeguarding user security and privacy remains imperative, particularly in view of the significant power these technologies exert.[75]

Another significant initiative promoted by the Holy See is the *Hiroshima AI Process Addendum on Generative AI*, a key document emerging from the Hiroshima AI Process launched by the G7 leaders and officially adopted on 30 October 2023.[76] Although not a legally binding text, the document – also signed by the Vatican – serves as a foundational reference for global AI governance. The *Addendum* emphasizes the need for ethical oversight of generative AI, reiterating the core principles advanced by the *Rome Call* and underscoring the imperative of developing AI that is inclusive, fair, and – given its profound social impact – reliable, safe, and privacy-preserving, so that its potential may be harnessed for the good of humanity.

---

72 Cf. Francis, *Address of His Holiness Pope Francis to Participants in the "Minerva Dialogues"*, Vatican City, 27.03.2023, URL: https://www.vatican.va/content/francesco/en/speeches/2023/march/documents/20230327-minerva-dialogues.html.

73 Francis, *Address Prepared by Pope Francis*, op. cit.

74 Generative AI systems can create coherent texts, but this does not ensure reliability. They may "hallucinate," producing statements that seem plausible but are false or biased. This is particularly dangerous in disinformation campaigns that undermine trust in the media. Privacy, data ownership, and intellectual property are also at risk. Misuse of these technologies can further lead to discrimination, electoral manipulation, mass surveillance, digital exclusion, and rising individualism detached from society. Cf. Francis, *Message of His Holiness Pope Francis for the 57th World Day of Peace: Artificial Intelligence and Peace*, op. cit., par. 4.

75 Cf. A. Adam, *Delegating and Distributing Morality: Can We Inscribe Privacy Protection in a Machine?*, "Ethics and Information Technology" 2005, Vol. 7, No. 4, pp. 233–242.

76 *Hiroshima Addendum*, URL: https://www.romecall.org/wp-content/uploads/2024/07/Hiroshima-Addendum-2.pdf.

In the search for an ethical framework for AI, the social doctrine of the Catholic Church reminds us that technologies must be studied and developed according to criteria that ensure their genuine service to the entire human family,[77] proposing an ethics of technological development grounded in the principles of human dignity, justice, subsidiarity, and solidarity.

The technology is not merely a tool but a complex force that requires careful ethical evaluation to ensure that it serves human dignity and the common good.[78] This common good is something towards which all people naturally aspire, and no ethical framework worthy of the name can fail to acknowledge it as a fundamental guiding principle.[79] It must therefore respond to the biblical mandate to "till and keep the earth" (Gen 2:15), strengthening the covenant between humanity and creation in accordance with God's creative love.[80] In the Catholic understanding, the human person possesses an irreducible spiritual transcendence that no machine or algorithm can replicate or replace. Only the human being, created "in the image and likeness of God" (Gen 1:27), has a spiritual and immortal soul, capable of moral discernment and free self-determination.

Technological development can contribute significantly to the progress of humanity, but it can also foster the illusion of human self-sufficiency when people focus solely on how to act, neglecting the deeper why that gives moral and spiritual meaning to their actions. However, such progress cannot truly benefit humanity unless it is accompanied by genuine moral and spiritual maturity: technological advancement, while representing a potentially great benefit for hu-

---

77 Francis, *Message of His Holiness Pope Francis to the Executive Chairman of the "World Economic Forum" on the Occasion of the Annual Gathering in Davos-Klosters*, 23–26.01.2018, URL: https://www.vatican.va/content/francesco/en/messages/pont-messages/2018/documents/papa-francesco_20180112_messaggio-davos2018.html.

78 For further discussion, see S.P. Chalmers, *Papal Teaching on the Ethical Challenges of Artificial Intelligence*, in: *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence*, eds. D.H. de la Iglesia, J.F. de Paz Santana, A.J. López Rivero, Springer, Cham 2023, pp. 167–177, https://doi.org/10.1007/978-3-031-14859-0_15.

79 Cf. Francis, *Address of His Holiness Pope Francis to the Participants in the Seminar "The Common Good in the Digital Age," Organized by the Dicastery for Promoting Integral Human Development (DPIHD) and the Pontifical Council for Culture (PCC)*, Vatican City, 27.09.2019, URL: https://www.vatican.va/content/francesco/en/speeches/2019/september/documents/papa-francesco_20190927_eradigitale.html.

80 Francis, *Laudato si'*: *Encyclical Letter on the Care for Our Common Home*, 24.05.2015, par. 109, URL: https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html.

mankind, must always be guided by an ethical conscience capable of discernment and responsibility.[81]

Given the contemporary context, there is an urgent need to ground the design, development, and use of AI in a robust ethical, anthropological, and wisdom-based foundation. It is necessary to overturn the assumption that everything technically possible is therefore legitimate, and instead ask how we can ensure that what is truly just becomes possible.[82] From this standpoint, the central challenge identified by the Catholic Church lies in orienting AI towards fostering a network of authentic communication – one rooted in communion that unites, in truth that sets free, and in love that confers ultimate meaning to human action.[83]

AI must always remain a tool at the service of humanity and must never replace human conscience or ethical discernment. Its orientation must consistently aim at the integral development of both the human person and society as a whole.[84] One of the critical concerns highlighted is the growing tendency towards the anthropomorphization of AI, which risks displacing authentic human relationships, particularly among younger generations. For this reason, the Church strongly emphasizes the necessity of education in critical thinking and discernment in the use of data and content generated by intelligent systems.[85]

Recently, Pope Leo XIV reaffirmed the Church's position on the development of AI, stressing that this epochal transformation requires careful reflection and ethical guidance to ensure its orientation towards humanity and the common good.[86] As AI systems acquire the capacity to make autonomous, technically driven decisions, it becomes imperative to examine their ethical and anthropo-

---

81 Cf. Benedict XVI, *Encyclical Letter Caritas in Veritate*, op. cit., par. 68–70.

82 Cf. Francis, *Address Prepared by Pope Francis*, op. cit.

83 Cf. Dicastero per la Communicazione, *La Chiesa di fronte all'attuale fenomeno dell'"intelligenza artificiale"*, 22.05.2024, URL: https://www.comunicazione.va/it/notizie/notizie_2024/la-chiesa-di-fronte-all-attuale-fenomeno-dell-intelligenza-artif.html.

84 Cf. Dicastery for the Doctrine of the Faith, Dicastery for Culture and Education, *Antiqua et nova: Note on the Relationship between Artificial Intelligence and Human Intelligence*, 28.01.2025, par. 6, URL: https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.htmlhttps://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html.

85 Cf. ibid., par. 21.

86 Cf. Leo XIV, *Message of the Holy Father, Signed by the Cardinal Secretary of State Pietro Parolin, on the Occasion of the AI for Good Summit 2025*, Geneva, 10.07.2025, URL: https://www.vatican.va/content/leo-xiv/en/messages/pont-messages/2025/documents/20250708-messaggio-aiforgood-ginevra.html.

logical implications. While AI may simulate human reasoning and perform tasks with remarkable efficiency, it remains incapable of exercising moral judgement or fostering authentic human relationships. For this reason, technological advancement must be accompanied by a strong commitment to human values, moral conscience, and a deepened sense of responsibility.[87] This unprecedented stage of innovation thus calls for renewed reflection on the meaning of human existence itself. Ultimately, AI requires ethical guidelines and regulatory frameworks grounded in the primacy of human dignity, rather than being governed solely by criteria of utility or efficiency.

The Church's moral and social teachings offer valuable guidance to ensure that AI is employed in ways that respect and preserve human agency. Reflections on justice, for instance, should also encompass the promotion of equitable social structures, the safeguarding of global security, and the advancement of peace. By exercising prudence, both individuals and communities can discern responsible ways to harness AI for the benefit of humanity, while avoiding applications that might compromise human dignity or cause harm to the environment.[88]

In contemporary debates on AI governance, the Catholic Church underscores the necessity of meaningful human oversight as an essential condition for orienting technological innovation towards the service of the human person and the common good.[89] This perspective does not remain at the level of abstract principles but identifies operational criteria capable of translating the values of human dignity, responsibility, and social justice into concrete regulatory practices.[90]

In light of the personalist principle and the categorical rejection of delegating life-or-death decisions to machines, meaningful human oversight in the military domain requires: (1) human-in-command structures with clearly identifiable legal responsibility across the entire chain of command; (2) *ex ante* limits on the autonomous functions of weapon systems,[91] excluding target selection and

[87] Cf. Francis, *Address of His Holiness Pope Francis to the Participants in the Seminar "The Common Good in the Digital Age"*, op. cit.

[88] Cf. Dicastery for the Doctrine of the Faith, Dicastery for Culture and Education, *Antiqua et nova*, op. cit., par. 47.

[89] Cf. *Catechism of the Catholic Church*, Libreria Editrice Vaticana, Vatican City 1992, par. 1905–1912, URL: https://www.vatican.va/archive/ENG0015/_INDEX.HTM.

[90] Cf. Francis, *Message for the 57th World Day of Peace: Artificial Intelligence and Peace*, op. cit., par. 2–10.

[91] In discussions on lethal autonomous weapon systems, Pope Francis made a pivotal statement at the 2024 G7 summit: "No machine should ever choose to take the life of a human being," affirming that decisions affecting life and death must remain under human authority. Cf. Francis, *Address*

engagement without effective human control; (3) compliance testing with international humanitarian law and human rights standards, including mandatory red-teaming and kill-switch mechanisms; and (4) full traceability through independent auditing and periodic review of rules of engagement.[92] To align AI with the principle of person-centred care and equity in access, the Catholic Church advocates for: (1) clinical governance of AI with ultimate medical responsibility remaining with the physician; (2) ethical-clinical impact assessments and post-market surveillance of devices and algorithms; (3) clinically useful explainability for both doctors and patients; (4) specific informed consent procedures for AI use, with safeguards for vulnerable groups; (5) systematic audits of bias and performance across diverse populations; and (6) robust data protection measures (minimization, quality, security), combined with human override mechanisms for inappropriate algorithmic recommendations.[93]

In accordance with the Catholic principles of the dignity of work and social justice,[94] meaningful human oversight in the field of employment must include: (1) human-in-the-loop mechanisms for adverse decisions (hiring, promotion, dismissal), guaranteeing the right to explanation and appeal; (2) impact assessments on non-discrimination and inclusion, with periodic audits and corrective measures; (3) participation of workers' representatives in the design and deployment of AI systems; (4) prohibition of black-box models for high-impact uses, accompanied by decision logs for accountability; and (5) continuous training on the critical use of algorithmic tools.

The Catholic perspective offers a coherent and universally applicable ethical framework that translates core values into concrete operational guidelines: prioritizing the human person, ensuring balance and prudence, advancing justice and inclusion, guaranteeing traceable accountability, and protecting the most vulnerable. These guidelines provide a foundational reference for both public policy and private regulatory practices, fostering a digital ecosystem genuinely oriented towards the common good.

---

*of His Holiness Pope Francis*, op. cit.; cf. A. Sharkey, *Autonomous Weapons Systems, Killer Robots and Human Dignity*, "Ethics and Information Technology" 2019, Vol. 21, No. 2, pp. 75–87.

92 Cf. Pontifical Academy for Life, *Rome Call for AI Ethics*, op. cit.

93 Cf. Francis, *Laudato si'*, op. cit., par. 109–110.

94 Cf. Pontifical Council for Justice and Peace, *Compendium of the Social Doctrine of the Church*, Libreria Editrice Vaticana, Vatican City 2004, par. 270–275, URL: https://www.vatican.va/roman_curia/pontifical_councils/justpeace/documents/rc_pc_justpeace_doc_20060526_compendio-dott-soc_en.html.

## 6. Conclusion

The profound transformations shaping contemporary society through the widespread use of AI inevitably raise significant ethical questions. Within this context, the Catholic contribution offers a coherent and universalizable framework of principles that does not remain purely theoretical but provides operational criteria: the centrality of the person, proportionality and precaution, justice and inclusion, traceable responsibility, and the protection of the most vulnerable.[95] Translated into practical requirements – such as effective human oversight, auditability, context-appropriate explainability, data protection, impact assessments, and redress mechanisms – these criteria are capable of informing both public and private standards and regulations, fostering a digital ecosystem genuinely oriented towards the common good.[96]

The Catholic Church's ethical evaluation of AI does not constitute a rejection of technological progress, but rather an appeal to orient innovation according to principles that safeguard human dignity and the common good.[97] The overarching goal is to ensure that AI remains at the service of humanity, promotes justice, and contributes to the construction of a more equitable, peaceful, and fraternal society.[98]

In this vision, human beings, endowed with their distinctive "wisdom of the heart," possess the capacity to discern the interconnectedness of realities, to recognize the positive dimensions of existence, and to uncover its deeper meaning.[99] This wisdom is neither reducible to abstract theory nor to mere technical expertise; rather, it is expressed concretely in relationships, commitment, and care.[100]

---

95 Cf. P. Benanti, *L'uomo è un algoritmo? Il senso dell'umano e l'intelligenza artificiale*, Castelvecchi, Roma 2025, pp. 45–48.

96 Cf. Francis, *Message for the 57th World Day of Peace: Artificial Intelligence and Peace*, op. cit., par. 2–6.

97 Cf. Francis, *Laudato si'*, op. cit., par. 102–114.

98 Cf. Francis, *Fratelli tutti: Encyclical Letter on Fraternity and Social Friendship*, 3.10.2020, par. 114–121, URL: https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20201003_enciclica-fratelli-tutti.html.

99 Cf. Francis, *Message of His Holiness Pope Francis for the 58th World Day of Social Communications: Artificial Intelligence and the Wisdom of the Heart. Towards a Fully Human Communication*, 24.01.2024, URL: https://www.vatican.va/content/francesco/en/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html.

100 Cf. V. Corrado, S. Pasta, eds., *Intelligenza artificiale e sapienza del cuore. Commento al Messaggio di Papa Francesco per la 58ma Giornata mondiale delle Comunicazioni Sociali*, Scholé, Brescia 2024, p. 102.

It enables the perception of realities that data alone cannot reveal, while recalling that at the foundation of all things lies the relational bond among persons – a dimension that digital technologies can neither replace nor diminish.

**Acknowledgement**
The author used generative AI tools exclusively for linguistic and editorial support, and their use did not have any significant impact on the scientific content of the article.

## Bibliography

Adam A., *Delegating and Distributing Morality: Can We Inscribe Privacy Protection in a Machine?*, "Ethics and Information Technology" 2005, Vol. 7, No. 4, pp. 233–242.

Atabekov A., Yastrebov A., *Legal Status of Artificial Intelligence across Countries: Legislation on the Move*, "European Research Studies Journal" 2018, Vol. 21, No. 4, pp. 773–782.

Beauchamp T.L., *Standing on Principles: Collected Essays*, Oxford University Press, New York 2010.

Beauchamp T.L., Childress J.F., *Principles of Biomedical Ethics*, Oxford University Press, Oxford 1979.

Beauchamp T.L., Childress J.F., *Principles of Biomedical Ethics*, 8th ed., Oxford University Press, New York–Oxford 2019.

Benanti P., *Homo Faber: The Techno-Human Condition*, EDB, 2018.

Benanti P., *L'uomo è un algoritmo? Il senso dell'umano e l'intelligenza artificiale*, Castelvecchi, Roma 2025.

Benanti P., *Oracoli. Tra algoretica e algocrazia*, Luca Sossella Editore, Roma 2018.

Benedict XVI, *Encyclical Letter Caritas in Veritate*, Rome, 29.06.2009, URL: https://www.vatican.va/content/benedict-xvi/en/encyclicals/documents/hf_ben-xvi_enc_20090629_caritas-in-veritate.html.

Berry D.M., *The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot*, "Weizenbaum Journal of the Digital Society" 2023, Vol. 3, No. 3, https://doi.org/10.34669/WI.WJDS/3.3.2.

Broussard M., *Artificial Unintelligence: How Computers Misunderstand the World*, The MIT Press, Cambridge, MA–London 2019.

*Catechism of the Catholic Church*, Libreria Editrice Vaticana, Vatican City 1992, URL: https://www.vatican.va/archive/ENG0015/_INDEX.HTM.

Cath C., *Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges*, "Philosophical Transactions of the Royal Society A" 2018, Volume 376, No. 2133, 20180080, https://doi.org/10.1098/rsta.2018.0080.

Chalmers S.P., *Papal Teaching on the Ethical Challenges of Artificial Intelligence*, in: *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence*, eds. D.H. de la Iglesia, J.F. de Paz Santana, A.J. López Rivero, Spinger, Cham 2023, pp. 167–177, https://doi.org/10.1007/978-3-031-14859-0_15.

Coeckelbergh M., *AI Ethics*, The MIT Press, Cambridge, MA, 2020.

Corrado V., Pasta S., eds., *Intelligenza artificiale e sapienza del cuore. Commento al Messaggio di Papa Francesco per la 58ma Giornata mondiale delle Comunicazioni Sociali*, Scholé, Brescia 2024.

Cucci G., *Emozioni e ragione: due mondi antitetici?*, "La Civiltà Cattolica" 2015, Vol. 3, pp. 139–150.

D'Avack, L., *La rivoluzione tecnologica e la nuova era digitale: problemi etici*, in: *Intelligenza Artificiale. Il diritto, i diritti, l'etica*, ed. U. Ruffolo, Giuffrè, Milano 2020, pp. 3–28.

Dicastero per la Communicazione, *La Chiesa di fronte all'attuale fenomeno dell'"intelligenza artificiale"*, 22.05.2024, URL: https://www.comunicazione.va/it/notizie/notizie_2024/la-chiesa-di-fronte-all-attuale-fenomeno-dell-intelligenza-artif.html.

Dicastery for the Doctrine of the Faith, Dicastery for Culture and Education, *Antiqua et nova: Note on the Relationship between Artificial Intelligence and Human Intelligence*, 28.01.2025, URL: https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html.

European Commission, *Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act)*, C(2025) 5052 final, 29.07.2025, URL: https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act.

European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, COM(2021)206final, 21.04.2021, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

European Commission, *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, 19.02.2020, URL: https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

European Union, *Regulation (EU) 2024/1689 of the European Parliament and of the Council*, 13.07.2024, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

Floridi L., *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford University Press, Oxford 2023.

Floridi L., *Soft Ethics, the Governance of the Digital and the General Data Protection Regulation*, "Philosophical Transactions of the Royal Society A" 2018, Vol. 376, No. 2133, https://doi.org/10.1098/rsta.2018.0081.

Floridi L., ed., *The Cambridge Handbook of Information and Computer Ethics*, Cambridge University Press, Cambridge 2010.

Floridi L., ed., *Ethics, Governance, and Policies in Artificial Intelligence*, Springer Verlag, Cham 2021.

Floridi L., Cabitza F., *Intelligenza artificiale. L'uso delle nuove macchine*, Bompiani, Firenze–Milano 2021.

Floridi L., Cowls J., *Unified Framework of Five Principles for AI in Society*, "Harvard Data Science Review" 2019, Vol. 1, pp. 2–15.

Floridi L., Cowls J., *A Unified Framework of Five Principles for AI in Society*, in: L. Floridi, ed., *Ethics, Governance, and Policies in Artificial Intelligence*, Springer Verlag, Cham 2021, pp. 5–17.

Floridi L., Sanders J.W., *Artificial Evil and the Foundation of Computer Ethics*, "Ethics and Information Technology" 2001, Vol. 3, No. 1, pp. 55–66.

Floridi L., Sanders J.W., *On the Morality of Artificial Agents*, "Minds & Machines" 2004, Vol. 14, No. 3, pp. 349–379.

Floridi L., et al., *AI 4 People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, "Minds & Machines" 2018, Vol. 28, No. 4, pp. 689–707.

Francis, *Address of His Holiness Pope Francis*, Borgo Egnazia, 14.06.2024, URL: https://www.vatican.va/content/francesco/en/speeches/2024/june/documents/20240614-g7-intelligenza-artificiale.html.

Francis, *Address of His Holiness Pope Francis to Participants in the "Minerva Dialogues"*, Vatican City, 27.03.2023, URL: https://www.vatican.va/content/francesco/en/speeches/2023/march/documents/20230327-minerva-dialogues.html.

Francis, *Address of His Holiness Pope Francis to the Participants in the Congress on "Child Dignity in the Digital World"*, Vatican City, 6.10.2017, URL: https://www.vatican.va/content/francesco/en/speeches/2017/october/documents/papa-francesco_20171006_congresso-childdignity-digitalworld.html.

Francis, *Address of His Holiness Pope Francis to the Participants in the Seminar "The Common Good in the Digital Age," Organized by the Dicastery for Promoting Integral Human Development (DPIHD) and the Pontifical Council for Culture (PCC)*, Vatican City, 27.09.2019, URL: https://www.vatican.va/content/francesco/en/speeches/2019/september/documents/papa-francesco_20190927_eradigitale.html.

Francis, *Address of Holy Father Francis*, Cagliari, 22.09.2013, URL: https://www.vatican.va/content/francesco/en/speeches/2013/september/documents/papa-francesco_20130922_cultura-cagliari.html.

Francis, *Address Prepared by Pope Francis, Read by H.E. Archbishop Paglia, President of the Pontifical Academy for Life, Meeting with the Participants in the Plenary Assembly of the Pontifical Academy for Life*, Vatican City, 28.02.2020, URL: https://www.vatican.va/content/francesco/en/speeches/2020/february/documents/papa-francesco_20200228_accademia-perlavita.html.

Francis, *Apostolic Exhortation Laudate Deum*, Rome, 4.10.2023, URL: https://www.vatican.va/content/francesco/en/apost_exhortations/documents/20231004-laudate-deum.html.

Francis, *Fratelli tutti: Encyclical Letter on Fraternity and Social Friendship*, 3.10.2020, URL: https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20201003_enciclica-fratelli-tutti.html.

Francis, *General Audience*, Saint Peter's Square, 5.06.2013, URL: https://www.vatican.va/content/francesco/en/audiences/2013/documents/papa-francesco_20130605_udienza-generale.html.

Francis, *Laudato si'*: *Encyclical Letter on the Care for Our Common Home*, 24.05.2015, URL: https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html.

Francis, *Letter of His Holiness Pope Francis to the President of the Pontifical Academy for Life for the 25th Anniversary of the Establishment of the Academy: Humana Communitas*, Vatican City, 6.01.2019, URL: https://www.vatican.va/content/francesco/en/letters/2019/documents/papa-francesco_20190106_lettera-accademia-vita.html.

Francis, *Message of His Holiness Pope Francis for the 57th World Day of Peace: Artificial Intelligence and Peace*, 1.01.2024, URL: https://www.vatican.va/content/francesco/en/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html.

Francis, *Message of His Holiness Pope Francis for the 58th World Day of Social Communications: Artificial Intelligence and the Wisdom of the Heart. Towards a Fully Human Communication*, 24.01.2024, URL: https://www.vatican.va/content/francesco/en/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html.

Francis, *Message of His Holiness Pope Francis to the Executive Chairman of the "World Economic Forum" on the Occasion of the Annual Gathering in Davos-Klosters*, 23–26.01.2018, URL: https://www.vatican.va/content/francesco/en/messages/pont-messages/2018/documents/papa-francesco_20180112_messaggio-davos2018.html.

Francis, *Message of Pope Francis for the 48th World Communications Day: Communication at the Service of an Authentic Culture of Encounter*, 1.06.2014, URL: https://www.vatican.va/content/francesco/en/messages/communications/documents/papa-francesco_20140124_messaggio-comunicazioni-sociali.html.

Hagendorff T., *The Ethics of AI Ethics: An Evaluation of Guidelines*, "Minds & Machines" 2020, Vol. 30, pp. 99–120.

Hajkowicz S., *Global Megatrends: Seven Patterns of Change Shaping Our Future*, CSIRO Publishing, Melbourne 2015.

High-Level Expert Group on AI (AI HLEG), *Ethics Guidelines for Trustworthy AI*, 8.04.2019, URL: https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf.

*Hiroshima Addendum*, URL: https://www.romecall.org/wp-content/uploads/2024/07/Hiroshima-Addendum-2.pdf.

Jobin A., Ienca M., Vayena E., *The Global Landscape of AI Ethics Guidelines*, "Nature Machine Intelligence" 2019, Vol. 9, No. 1, pp. 389–399.

Kroll J.A., et al., *Accountable Algorithms*, "University of Pennsylvania Law Review" 2017, Vol. 165, No. 3, pp. 633–705.

Latonero M., *Governing Artificial Intelligence: Upholding Human Rights & Dignity*, Data & Society, URL: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Leo XIV, *Message of the Holy Father, Signed by the Cardinal Secretary of State Pietro Parolin, on the Occasion of the AI for Good Summit 2025*, Geneva, 10.07.2025, URL: https://www.vatican.va/content/leo-xiv/en/messages/pont-messages/2025/documents/20250708-messaggio-aiforgood-ginevra.html.

Lin P., Abney K., Bekey G., *Robot Ethics: Mapping the Issues for a Mechanized World*, "Artificial Intelligence" 2011, Vol. 175, Nos. 5–6, pp. 942–949, https://doi.org/10.1016/j.artint.2010.11.026.

Morley J., Floridi L., Kinsey L., Elhalal A., *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" 2020, Vol. 26, pp. 2141–2168, https://doi.org/10.1007/s11948-019-00165-5.

Morley J., et al., *Ethics as a Service: A Pragmatic Operationalisation of AI Ethics*, "Minds & Machines" 2021, Vol. 31, No. 2, pp. 239–356, https://doi.org/10.1007/s11023-021-09563-w.

Müller V.C., *Ethics of Artificial Intelligence and Robotics*, in: *Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), ed. E.N. Zalta, URL: https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/.

Pontifical Academy for Life, *Rome Call for AI Ethics*, Rome, 28.02.2020, URL: https://www.vatican.va/roman_curia/pontifical_academies/acdlife/documents/rc_pont-acd_life_doc_20202228_rome-call-for-ai-ethics_en.pdf

Pontifical Council for Justice and Peace, *Compendium of the Social Doctrine of the Church*, Libreria Editrice Vaticana, Vatican City 2004, URL: https://www.vatican.va/roman_curia/pontifical_councils/justpeace/documents/rc_pc_justpeace_doc_20060526_compendio-dott-soc_en.html.

Quintarelli S., Corea F., Fossa F., Loreggia A., Sapienza S., *AI: profili etici. Una prospettiva etica sull'Intelligenza Artificiale. Princìpi, diritti e raccomandazioni*, "BioLaw Journal – Rivista di BioDiritto" 2019, Vol. 3, pp. 183–204.

RenAIssance Foundation, *Call for AI Ethics*, 28.02.2020, URL: https://www.romecall.org/.

Sadin E., *Critica della ragione artificiale. Una difesa dell'umanità*, Luiss University Press, Milano 2019.

Sharkey A., *Autonomous Weapons Systems, Killer Robots and Human Dignity*, "Ethics and Information Technology" 2019, Vol. 21, No. 2, pp. 75–87.

Smart Dubai, *AI Ethics Principles & Guidelines*, 2018, URL: https://www.digitaldubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf.

Taddeo M., Floridi L., *How AI Can Be a Force for Good: An Ethical Framework Will Help to Harness the Potential of AI while Keeping Humans in Control*, "Science" 2018, Vol. 361, No. 6404, pp. 751–752.

UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, Paris 2022, URL: https://unesdoc.unesco.org/in/documentViewer.xhtml?v=2.1.196&id=p::usmarcdef_0000381137&file=/in/rest/annotationSVC/DownloadWatermarkedAttachment/attach_import_75c9fb6b-92a6-4982-b772-79f540c9fc39%3F_%3D381137eng.pdf&locale=en&multi=true&ark=/ark:/48223/pf0000381137/PDF/381137eng.pdf#1517_21_EN_SHS_int.indd%3A.8946%3A.

United Nations, *Universal Declaration of Human Rights*, URL: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

Weizenbaum J., *Il potere del computer e la ragione umana. I limiti dell'intelligenza artificiale*, EGA-Edizioni Gruppo Abele, Torino 1987.

Winner L., *Do Artifacts Have Politics?*, in: L. Winner, *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press, Chicago 1988, pp. 19–39.

# Notes about the Authors

Brian Ball – dr, Northeastern University London, Devon House, 58 St Katharine's Way, London E1W 1LP, United Kingdom, brian.ball@nulondon.ac.uk, ORCID: 0000-0003-2478-6151.

Alex Cline – dr, Queen Mary University of London, London City Institute of Technology, 5 Hope St, Leamouth Peninsula, London E14 0BZ, United Kingdom, a.cline@qmul.ac.uk, ORCID: 0000-0002-7723-7175.

David Peter Wallis Freeborn – dr, Northeastern University London, 58 St Katharine's Way, London E1W 1LP, United Kingdom, david.freeborn@nulondon.ac.uk, ORCID: 0000-0002-2117-8145.

Alice C. Helliwell – dr, Northeastern University London, 58 St Katharine's Way, London E1W 1LP, United Kingdom, alice.helliwell@nulondon.ac.uk, ORCID: 0000-0003-0147-7700.

Sara Lumbreras – dr, Universidad Pontificia Comillas, Instituto de Investigación Tecnológica, Calle del Rey Francisco, 4 – 28008 Madrid, España, slumbreras@comillas.edu, ORCID: 0000-0002-5506-9027.

Max Parks – dr, University of Michigan, Mott Community College, 1401 E Court St, Flint, MI 48503, United States, maxaeonparks@gmail.com, ORCID: 0009-0003-5278-6965.

Luka Perušić – dr, University of Zagreb, Faculty of Humanities and Social Sciences, Ivana Lučića 3, 10000, Zagreb, Croatia, lperusic@yahoo.com, ORCID: 0000-0002-5339-781X.

Neomal Silva – mgr, Melbourne, Australia, neomals@yahoo.co.uk, ORCID: 0009-0001-8886-6620.

Krzysztof Trębski – dr, Trnava University in Trnava, Faculty of Theology, Department of Philosophy, Kostolná 1, P.O.BOX 173, 814 99 Bratislava, Slovakia, kris.treb@gmail.com, ORCID: 0000-0003-0115-5787.

Andrea Vestrucci – prof., Universität Bamberg, Department of Computer Science, Kapuzinerstraße 16, D-96047 Bamberg, Germany, andrea.vestrucci@uni-bamberg.de, ORCID: 0000-0002-6336-1036.

Aleksandra Vučković – dr, University of Belgrade, Faculty of Philosophy, Institute for Philosophy, Čika-Ljubina 18-20, Belgrad 11000, Serbia, aleksandra.vuchkovic@gmail.com, ORCID: 0000-0002-5960-2909.

Ralph Stefan Weir – dr, University of Lincoln, School of Humanities and Heritage, Brayford Pool, Lincoln LN6 7TS, United Kingdom, RWeir@lincoln.ac.uk, ORCID: 0000-0001-7159-3497.