Journal of Ethics and Emerging Technologies

IEET

Book Review

Review of *Ethics of Artificial Intelligence*

Ethics of Artificial Intelligence. Eds. Francisco Lara, Jan Deckers. Springer, 2023.

Pietro Intropi[1]*

[1] MSCA post-doctoral fellow; Université Catholique de Louvain, Chaire Hoover d'étique économique et sociale (ISP), Centre Charles de Visscher pour le droit international et européen (CeDIE); pietro.intropi@uclouvain.be

* Correspondence: pietro.intropi05@gmail.com

Abstract: In this review, I summarise and briefly discuss the themes of the book, suggesting two interpretative perspectives: a political philosophy one that highlights the risks that AI creates for society and the need for regulation, a meta-ethics/normative one that focuses on how the AI revolution challenges traditional interpretation of moral concepts, like moral status, autonomy, responsibility, and authorship.

Keywords: AI ethics, applied ethics, regulation, moral status, machine discrimination

Citation: Intropi, Pietro. 2026. Review of Ethics of Artificial Intelligence. *Journal of Ethics and Emerging Technologies* 36: 1. https://doi.org/10.55613/jeet.v36i1.235

Received: 27/04/2026
Accepted: 28/04/2026
Published: 30/04/2026

Publisher's Note: IEET stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The literature on AI ethics is blooming. Yet, as the authors notice in the Introduction (p. 2), there is only a limited number of edited collections that present the main debates and most salient questions within this emerging field of philosophical inquiry (see e.g. Edmonds 2024). This is why the book edited by Francisco Lara (Granada University) and Jan Deckers (Newcastle University) should be welcomed as a valuable contribution: it features essays written mostly by academics from Spanish universities, and it can be read as a general introduction to AI ethics, or used to structure a reading list for a course in moral/political philosophy. In this review I introduce and briefly discuss the themes of the book suggesting two interpretive perspectives: (1) a political philosophy one that highlights the risks that AI creates for society and the need for regulation, and (2) a meta-ethical/ normative ethics one that focuses on how the AI revolution challenges traditional interpretations of moral concepts like moral status, autonomy, responsibility, authorship.

The book contains an introduction and 11 chapters. The range of topics is quite broad, and includes: machine discrimination (ch. 2), the challenges of building a trustworthy and explainable AI (versus the opacity of AI systems) (ch. 3), AI and moral status (ch. 4). These are followed by applied chapters that discuss: virtual assistants (ch. 5), virtual reality (ch. 6), deepfakes in the artistic domain (ch. 7), care robots (ch. 8), autonomous weapon systems (ch. 9). The final part of the book (Part III) focuses on: AI governance – with a particular focus on the EU (ch. 10), AI and sustainability (ch. 11), singularity and mind-uploading (ch. 12). There are many points of contact between the chapters. For instance, inquiring about how responsibility, agency, and autonomy can be attributed to AI entities (ch. 4) is also relevant to address ethical questions concerning human control on

autonomous weapon systems (ch. 9). The fact that some important topics are left out is perhaps an inescapable feature of edited collections. Nevertheless, it is surprising to see that there is no dedicated chapter on the impact of AI on labour markets, given how pressing are the effects of AI through job automation (though see p. 97 for some thoughts on the issue).

The ethical questions discussed in the book are inevitably intertwined with socio-political worries that concern the regulation of AI systems in democratic societies. Chapter 10 ("Ethical Principles and Governance for AI") is a good entry point for this politics-oriented perspective. Given how serious are the potential risks associated with the uses and misuses of AI, the need for regulations is undeniable. Yet, who should be responsible for issuing binding rules, and exactly what should be the function and content of such rules remain open questions. Ch. 10 discusses how private actors (e.g. companies) and public agencies (e.g. the state, the EU) can and should play a role in creating a framework for AI governance. We shouldn't rely only on private companies to entirely self-regulate: they may focus only on a subset of the relevant risks for society, and in any case the coercive authority of a public agent is required to enforce full compliance (see pp. 195-200).[1] So, a public system of rules is needed (p. 200). The chapter focuses especially on the EU, which is said to represent "the best example of a democratic governance of AI" (p. 195). The chapter was written whilst the EU AI act was under discussion, so some points may need updating. The chapter provides an insightful taxonomy for thinking about AI governance at different levels (company-level, state-level, EU-level, and globally).[2]

Some of the most pressing risks and challenges that AI creates for society concern: machine discrimination, the excessive delegation of human decisions to AI systems, human dependence on AI, alienation, deskilling, human obsolescence, privacy violations, the environmental costs of AI, systemic risks concerning the end of human civilization. Chapter 2 deals with the issue of machine discrimination: the author (Jorge Casillas) emphasizes that in delegating decisions to machines we risk "forgetting along the way that machines also fail, and the consequences are a thousand times more serious" (p. 14). For one, because they are "massive", possibly adversely affecting a great number of people, due to the fact that machines can take "millions of decisions in a second" (p. 15). Machine discrimination – where some groups are unfairly treated due to biases/flaws of machine learning models – can happen in a variety of domains: hiring, sentencing, access to loans etc. There are different kinds of biases/flaws that can skew machine decision-making (pp. 29-31), and the chapter suggests ways of avoiding/reducing them in different phases of the decision-making process (pp. 31-35). Hence, the need for a trustworthy AI, which is the

[1] That said, the adoption of codes of conduct by private companies is an important tool to meet objectives of social corporate social responsibility (p. 197).

[2] On the EU AI Act, see : https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai; https://www.theguardian.com/world/2023/dec/08/eu-agrees-historic-deal-with-worlds-first-laws-to-regulate-ai

issue taken up in chapter 3. AI systems are often not transparent, in the sense that the machine learning model that operationalises them is a black box: we know the initial data and the final outputs, but we don't really entirely understand the reasoning behind the processing (which relies on a series of very complicated mathematical operations) (p.43). So, increasing the transparency and interpretability of machine learning models is morally required. At points, this chapter becomes very technical, which is perhaps not entirely suitable for an edited collection of this type.

The discussion of applications of AI to specific domains is also very informative (Part II of the book). The empowerment effect of AI is undeniable: AI can assist us with a great number of tasks. Take the case of virtual assistants (e.g. Siri, Alexa) or of care robots, discussed in ch. 5 and ch. 8 respectively. We can rely on AI systems (e.g. ChatGPT) to correct our texts, extrapolate information, or help us finding the road to a place, but perhaps at the cost of becoming increasingly dependent on these assistive technologies and of eroding our cognitive abilities (deskilling): will this ultimately lead to the practical irrelevance of human skills (human obsolescence) (pp. 94-98)? [3] The socio-political implications are evident: think of how AI is revolutionizing professions and labour markets, and the impact of AI on education. This chapter invites us to think about what is really valuable about having *human beings* performing certain tasks (as opposed to an AI system). One answer can be suggested by reading ch. 8, where the authors discuss assistive technologies in the care domain. Care robots can help with "taking (…) medications, cooking, cleaning, and moving around" as well as with providing "emotional and social assistance" (pp. 151-152). Yet, at present they cannot "car(e) holistically", but only discharge specific caring tasks (p. 153). This prompts us to reflect about whether caring has an ineliminable human component (see p. 152), which is not to deny that a proper employment of assistive technologies can be very beneficial: e.g. it can relieve care-givers from burdensome tasks, allowing them more free time (p. 152).

Although it cannot be said that the chapters of the book explicitly share a common methodological approach, the book inspires the reader to adopt a 'grounded' perspective on AI. The issues that more urgently require philosophical attention are those that affect our daily lives, now and in the foreseeable future. Of course, we should also worry about the existential risks for humanity that would be brought about by building a kind of Artificial General Superintelligence that would exponentially surpass human intelligence (ch. 12). But, as we learn from ch. 12, such philosophical inquiries should be put in context: for example, as the chapter suggests, we should have a sound empirical discussion of what are the realistic possibilities of ever getting to the Singularity stage, contributing to deflate some of the futuristic hype that we find in these debates. AI's environmental costs ad impact on sustainability should really be a matter of concern, especially given that general public is insufficiently aware of them. Ch. 11 provides the analytical tools for reflecting

[3] A third source of harm discussed in the chapter is that of potential privacy violations (see Véliz 2020).

about this, inviting us (1) to think about the carbon-footprint of AI in all the phases of an AI system lifecycle ("the manufacturing of hardware, material extraction, the building of the systems" data storage, etc. p. 222) (pp. 221-223), and (2) to reflect on which kind of ethical theory is more apt to capture our sustainability-related obligations. To what extent should such obligations take into account how AI harms not only humans, but also non-human species, and the ecosystem as a whole (pp. 225-231)?

AI not only impacts on social lives, but also on how we reflect about the human experience: for example, on how we theorise core ethical ideas like responsibility, authorship, moral status. Take responsibility: imagine that an assistive robot makes a life-threatening mistake in giving medications to a vulnerable patient. Who should be held responsible: the robot or the humans who created the robot (see p. 161 and pp. 67-71)? Which kinds of responsibility are at stake? More generally, how should we re-think responsibility practices in light of the fact that in a foreseeable future we will be increasingly interact with *agential* AI entities? How should we conceptualise the kind of agency that AI entities seem to display? Is it of the kind that suffices for attributions of moral status? Do machines need to possess consciousness to be ascribed moral status? Ch. 4 summarizes the main positions within the debate on AI and moral status, which is of course relevant for thinking about whether AI entities can have rights, legal personality, and intrinsic value. In keeping with the grounded approach emphasized above, consider the case of autonomous weapon systems (ch. 9). Warfare technologies that will be used or are already being used could autonomously select the targets once programmed. Hence, it seems that the definition of autonomy that is apt in this context is one that allows to distinguish between different degrees of supervision and control that humans can exercise over the system ("human in/on/out of the loop", p. 173), rather than one that, say, speculates about whether weapon systems are conscious (ch. 9, see p. 171).

The book also invites us to reflect on other philosophical fault lines: for example, that between true and false, real and not real. Ch. 6 discusses virtual reality, which are immersive and interactive experiences rendered possible by developments in computer technology (p. 109-110). A key question is whether they can be classified as 'real' in the same sense in which the external physical world is deemed to be real (pp. 111-112). The chapter also illustrates the benefits and harms associated with virtual reality: on the one hand, the enormous expansion of opportunities to live (virtual) experiences of any kind; on the other hand, the psychological, health, and social risks – e.g. social isolation, "depersonalization (feelings of alienation towards one's own self) and derealization (feeling of detachment from reality) disorder" (p. 118) (see pp. 115-119). Ch. 7 also offers an interesting perspective on the blurring line between real/non-real, discussing ethical issues raised by deep-fakes in the specific domain of artistic production. The possibility of creating cultural items that closely follow stylistic patterns of renowned artistic figures (or that result from hybridizing styles) not only makes us thinking about whether AI-generated content can be classified as art – in such a way that the meaning of Art becomes

ever more contentious – but also make us revisiting traditional understandings of authorship (see pp. 139-145). For instance, how are we to classify the products of an AI tool that creates "a painting in the style of Rembrandt or a piece of music in the style of Bach" (p. 137)? Are these artistic objects? There was a time when making tools was understood to be a human prerogative, before we discovered that also great apes do it. Analogously, we may have to rethink the notion of authorship and the boundaries between art and craftmanship. This book offers a multifaceted picture of the ways in which AI is revolutionizing our personal lives and communal institutions, making it a very enjoyable and rich introduction to AI ethics.

References

(Edmonds (ed) (2024) Edmonds, David. 2024. *AI Morality*. Oxford: Oxford University Press.
(Véliz 2020) Véliz, Carissa. 2020. *Privacy Is Power*. London: Penguin Books.