Book review: Marc Coeckelbergh, AI Ethics, Mit Press, 2021

Ethics of AI: The Philosophical Challenges

Santoni de Sio, Filippo

DOI
10.1007/s11948-021-00323-8

Publication date
2021

Document Version
Final published version

Published in
Science and Engineering Ethics

Science and Engineering Ethics (2021) 27:50
https://doi.org/10.1007/s11948-021-00323-8

BOOK REVIEW

Mark Coeckelbergh, *AI Ethics*, Mit Press, 2021

Ethics of AI: The Philosophical Challenges

Filippo Santoni de Sio[1]

Received: 25 May 2021 / Accepted: 8 June 2021 / Published online: 2 August 2021

Everybody nowadays talks about ethical and societal issues with artificial intelligence, or, in short: AI Ethics. How to control the spreading of fake news on social networks? How to prevent facial recognition technology to foster discrimination and violation of civil rights? Should we ban fully autonomous lethal weapon systems? Who is responsible for lethal traffic accidents involving automated driving systems? Is it fair to be assessed by an automated system in a job or administrative procedure? However, debates in the media often lack philosophical breadth, depth and perspective. Mark Coeckelbergh has worked in the field of ethics and philosophy of technology for many years now. His recent short book AI Ethics aims to give a philosophical perspective on AI Ethics.

His perspective is philosophical in many good ways. First, from a methodological point of view, he starts by defining the problems of AI Ethics and putting them in the broader context of the history of ideas, often taking a refreshing critical stance towards many mainstream narratives about AI Ethics, narratives propagated by media, business—and bad philosophy (e.g. AI taking control over humanity, AI as the panacea of all societal problems). Second, he maps the current debates in AI Ethics to some older debates in philosophy of mind, philosophy of science and technology, and metaphysics (chapters 1–6). Third, he connects some of the current issues in AI Ethics with some broader topics in moral, legal and political philosophy (chapters 7–9). Fourth, also based on his first-hand experience in some policy initiatives on AI Ethics, he offers a critical presentation of some existing policy documents and some open challenges for the realization of a (global) policy over AI (chapters 10–11). Finally, he reflects on the position of AI Ethics in relation to the other big global challenges of the twenty-first century, in particular the environmental emergency (chapter 12).

The first, philosophical, part of the book is a pleasure to read, and is the one where the author seems more comfortable in striking a good balance between the

✉ Filippo Santoni de Sio
f.santonidesio@tudelft.nl

[1] Section Ethics/Philosophy of Technology, Delft University of Technology, Delft, The Netherlands

informal discussion of big questions and the presentation of some related philosophical theories. Particularly good is the discussion of the alleged risks of superintelligence and singularity: AI becoming so powerful to outsmart humanity and thereby becoming a threat to its existence; and the related idea of transhumanism: human beings encouraged to go beyond their limits to face the future. As Coeckleberg argues, these concerns are not only ethically irrelevant for now, as we don't seem close to anything like so-called general artificial intelligence, but also not particularly original from a philosophical point of view. Similar concerns had already been raised, for instance, by John von Neumann in the 1950s. Broadening the picture, Coeckelbergh gives a beautiful overview of different stories involving dreams and fears of human miraculous creations: from the myth of Pigmalion anticipating the (male) dream of female "sex robots", to the stories of the Golem and Frankenstein, exemplifying the fear of technology going out of control, to the Epic of Gilgamesh, an iconic representation of the eternal human dream of becoming immortal. In another insightful passage, he also connects the dreams of AI "abstracting pure forms from the messy material world" to the Platonic and Gnostic aspiration to liberate forms and models from the world of appearances.

The first part of the book also contains a discussion of the origin of the philosophical debate on the relationship between human and artificial intelligence, which Coeckelbergh traces back to Hubert Dreyfus' book *What Computers Can't Do* (1972). Dreyfus heavily criticized the possibility of artificial intelligence based on a phenomenological approach to knowledge. Interestingly, similar arguments would be adopted in the successive decades by philosophers coming from a more analytic tradition such as John Searle, and even by AI researchers themselves, under the name of embodied and situated cognition.

After a chapter on the debate on the moral status of AI, a chapter clarifying the different forms of AI—from symbolic AI to Machine Learning and beyond—and a chapter on the key role of data science for AI (Ethics), the book moves to a second part, on present-day ethical issues with AI. Coeckelbergh manages to present many important ethical topics in the space of three chapters, even though, especially if compared to the first part, it feels like the topics in these chapters are a bit compressed, their succession not always smooth and their distribution unconventional and sometimes slightly disorienting.

For instance, chapter 7 titled "Privacy and the other usual suspects" does not really present the debates on privacy and the major political relevance it has recently gained. The big political power of big tech like Google and Facebook is grounded in the massive acquisition and use of personal data of users. The manipulative and exploitative policies of these big techs are not only ethically problematic in itself, but also a symptom of a deeper problem. These companies are not subject to any effective economic, legal, political, social control. The book mentions the problem of accumulation of power in the hands of few actors in the later chapters on policy, but misses the opportunity to discuss the origin of this power. This origin has to be found in the design of the socio-technical infrastructures on which data are acquired, and their monopolistic character. From a philosophical point of view, one might have expected at least some reference to classic philosophical themes, such as the concepts of the "panopticon" (Foucault) and "the societies of control" (Deleuze).

Also, it is surprising to not find a broader discussion of the philosophical and legal challenges to redefine privacy and data protection in the age of big data and AI.

The chapter does have the merit to mention less known but very important issues, such as the ethically problematic exploitation of the so-called free "digital labour"—users producing the data and information over which tech companies produce their profit—and the challenge of protecting more vulnerable users from various forms of technologically mediated exploitation and manipulation: children, elderly persons, people with mental disabilities. Yet it would have been good to find some more explicit reference to philosophical concepts that might help making sense of- and address these issues, for instance the idea of human mental work as a form of capital to be protected, promoted and rewarded (Naastepad and Mulder 2018), or the idea of applying Amartya Sen's capability approach to digital technology (Oosterlaken 2013).

Chapter 8 on responsibility offers an explanation of why attributions of responsibility to human actors may become more and more problematic in the age of AI. Following Aristotle, Coeckelbergh refers to the so-called control and epistemic conditions for moral responsibility. In a nutshell: people should be held responsible only for the things that depend on their intentional choices and actions (control condition) and only if they are in the position to know what they are doing (epistemic condition). AI seem to create issues for both of these conditions. Can I really be blamed for causing an accident with my (partially) automated car, if wasn't able to properly interact with it, and had no idea that it could behave in a certain way under some circumstances? In addition, AI may aggravate the old "problem of many hands" in complex organisations—how to attribute individual responsibility for the outcome of process of interaction of many different agents. It does so by adding "many (intelligent) things" to the already existing many human hands. Bureaucrats famously tried to escape their responsibilities in Nuremberg by claiming to be part of a big organization they couldn't control. Are we facing a digital Nuremberg, where professionals and laypersons can too often appeal to the complexity of the socio-technical system of which they are part—in public administration, information, education, work, warfare etc.—as a way to avoid their accountability? This leads to the discussion of the problem of lack of transparency and explainability of processes involving AI (machine learning), important for accountability. Coeckelbergh rightfully points out that explainability is not only a technical issue solvable, for instance, by more transparent or so-called explainable AI. It is also a societal challenge—how can ordinary people get access to these explanations? And it also contains a philosophical challenge: defining what kind of explanation is needed in different contexts, and, ultimately, what counts as a good explanation for the behaviour of AI systems that include artificial agents. Issues of moral responsibility and accountability are clearly presented in this chapter, even though, here again, one might have expected some more explicit reference to some philosophical concepts, such as the problem of many hands (Thompson 1980; Van de Poel et al. 2015) and the so-called "responsibility gap" with AI (Matthias 2004).

Chapter 9 ("Bias and meaning of life") is centered on issues of (social) justice. The first part presents the issue of AI bias. Far from being a neutral tool, artificial intelligence tends to perpetuate and possibly aggravate the biases of the human

actors who design it and provide the data for their functioning. What is wrong with it, why can machines not just reflect our world full of biases, as opposed to make it (morally) better? At what point do biases become discriminatory, or otherwise unacceptable? What this question shows, Coeckelbergh argues, is that we need a theory of justice to make sense of- and address the concern about AI bias. Again, some more pointers into theories and concepts that might help shaping this reflection on justice would have been helpful: to what extent would traditional theories of civil rights suffice? Should we look more at theories of distributive justice à la Rawls, critical theories à la Habermas, feminist theories? The second part addresses the concern that AI may create massive unemployment, and asks whether this would really be bad news. Can't we find other, possibly more meaningful, ways to make our living, for instance by giving financial rewarding to social or care activities? Can Universal Basic Income be a solution? Questions about the meaning of work and life are here intertwined to political and economic issues of justice: who will be able to really enjoy a life without traditional paid work, and, we may add, shouldn't we rather pro-actively govern the transition to AI-mediated work in a more inclusive and democratic way rather than just accepting it as a destiny whose negative effects we should try to mitigate?

The chapter, the last in the ethics part, closes with a very important methodological question: to what extent may old philosophical concepts, such as for instance an Aristotelian virtue ethics, be of help in addressing new challenges of AI Ethics. Coeckelbergh suggests that they may be, even though he recognizes two big challenges: the speed at which both technology and society change doesn't seem to leave sufficient space for developing old individual virtues such as practical wisdom, and the plurality of cultures in society seem to require a more complex and pluralistic ethical approach. One may also add that quite simply new technological and societal challenges also require new concepts. Coeckelbergh rightfully suggests to also look outside the Western tradition to find new inspirations. However, the book could have been enriched by some more suggestion for new concepts and theories within the Western tradition to guide a future AI Ethics. Might we, for instance, need new concepts of privacy (Taylor et al. 2017), capital (Naastepad and Mulder 2018), democracy (Bernholz et al. 2021), control (Santoni de Sio and van den Hoven 2018), justice (Nagenborg 2009), responsibility (Santoni de Sio and Mecacci 2021)…? And, from a more methodological point of view, if current concepts of Responsible Innovation (Stilgoe et al. 2013) and Design for Values (van den Hoven et al. 2015) have shown some limitations, how can they be revised and integrated (Blok 2014; Kiran 2012)?

Chapters 10 and 11, which form a third part of the book, offer a critical presentation of some existing policy proposals in the field of AI Ethics. They provide a good overview of existing policy documents, not only those originating from the US and EU, and some European national state initiatives such as France and Austria but also, very importantly, China. There is also a brief overview of possible legal responses, in particular various forms of liability regimes, and of the debate on legal personhood for AI and Robots. An interesting, but not so surprising, observation is that "AI ethics policy are remarkably similar". Interesting because one would expect more diversity in proposals coming from different fora

(academic, political, industry) and different countries, but at the same time not so surprising. In fact, what makes the difference is the interpretation and operationalization of these principles in relation to concrete technologies in concrete social contexts: "it is one thing to name a number of ethical principles and quite another to figure out how to implement them in practice." (165). Philosophical analysis is very important here as well, since "the definition will shape the measures one proposes". For instance, Coeckelbergh mentions the global political convergence on the idea of a ban of "fully autonomous lethal weapon systems" as a rare case where an ethical campaign (stopkillerrobots.org) has reached a tangible political result. However, one may argue that this is precisely a case where the political result may be illusory, if we do not agree on the specific content of the prohibition: What "full autonomy" means, and what kind of human control is needed in order for the system to be (legally) acceptable (Amoroso and Tamburrini 2019).

Coeckelbergh also mentions the importance of more interdisciplinary and transdisciplinary collaboration on the topic of AI Ethics, and new forms of education mixing science, engineering and the humanities. Given his belief in the importance of developing virtues and his long experience as researcher and philosophy lecturer within a technical university (Twente), one would have expected some more emphasis on the importance of engineering education for the future of AI Ethics. Better research and education, however, are far from sufficient, as AI Ethics is also and above all a big political and economic challenge: if technological and data power is more and more concentrated in a few hands, it is unlikely that any concrete action to realise ethical AI will be possible. In this sense, Coeckelbergh admits, recent history does not give many reasons to be optimistic: looking at the slowness and ineffectiveness of global politics to address the environmental crisis makes one wonder why we should believe that things will be different in relation to the governance of AI.

In fact, the final chapter of the book even wonders if we should bother so much about the emerging issues of AI Ethics when other allegedly more urgent global emergencies are still far from being addressed: raising inequalities, wars, poverty, lack of access to water, migrations, climate change… Consistently with its overall approach, the book closes with a beautiful philosophical note. Is AI (and technology more generally) part of the problem or rather part of the solution? AI can help address some human and societal problems, provided it is seen for what it is: a human creation that is part of human life, not a transcendent, almost divine force that will lift us from our condition. We should not follow the prophets of transhumanism in their non-existing future worlds, or Elon Musk and other techno-prophets in their space escapes from earth. Life is here and now. The book ends where it started. In one of the first chapters, Coeckelbergh criticized the superintelligence and transhumanist approaches by comparing them to the Platonic dream of freeing humanity from their bodily limitations to elevate to the realm of pure, abstract forms. The last chapter mentions Hannah Arendt's warning about the risk of scientific and technological dreams abstracting and alienating from our human condition. AI can help us, only if it doesn't become an "alienation machine, an instrument to leave the Earth and deny our vulnerable, bodily, earthly, and dependent existential condition."

References

Amoroso, D., & Tamburrini, G. (2019). What Makes Human control Over Weapon Systems "Meaningful"? *ICRAC Working Paper Series #4*. https://www.icrac.net/wp-content/uploads/2019/08/Amoroso-Tamburrini_Human-Control_ICRAC-WP4.pdf

Bernholz, L., Landemore, H., & Reich, R. (2021). *Digital technology and democratic theory*. Chicago: Chicago Press.

Blok, V. (2014). Look who's talking: Responsible innovation, the paradox of dialogue and the voice of the other in communication and negotiation processes. *Journal of Responsible Innovation, 1*(2), 171–190. https://doi.org/10.1080/23299460.2014.924239

Kiran, A. H. (2012). Does responsible innovation presuppose design instrumentalism? Examining the case of telecare at home in the Netherlands. *Technology in Society, 34*(3), 216–226. https://doi.org/10.1016/j.techsoc.2012.07.001

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

Naastepad, C. W. M., & Mulder, J. M. (2018). Robots and us : Towards an economics of the ' Good Life.' *Review of Social Economy, 6764*, 1–33. https://doi.org/10.1080/00346764.2018.1432884

Nagenborg, M. (2009). Designing spheres of informational justice. *Ethics and Information Technology, 11*(3), 175–179. https://doi.org/10.1007/s10676-009-9200-3

Oosterlaken, I. (2013). *Taking a capability approach to technology and its design: A philosophical exploration*. Delft: Delft University of Technology.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy and Technology*. https://doi.org/10.1007/s13347-021-00450-x

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and A, I*, 5. https://doi.org/10.3389/frobt.2018.00015

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy, 42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Taylor, L., Floridi, L., & van der Sloot, B. (2017). *Group privacy: New challenges of data technologies*. Berlin: Springer.

Thompson, D. F. (1980). Moral responsibility of public officials: The problem of many hands. *The American Political Science Review, 74*(4), 905–916.

Van de Poel, I., Royakkers, L. M. M., & Zwart, S. D. (2015). *Moral responsibility and the problem of many hands*. Oxfordshire: Routledge. https://doi.org/10.4324/9781315734217

Van den Hoven, J., Vermaas, P. E., & van de Poel, I. (2015). *Handbook of ethics, values, and technological design: Sources, theory, values and application domains*. Springer. https://doi.org/10.1007/978-94-007-6970-0