AI & SOCIETY (2021) 36:661–664
https://doi.org/10.1007/s00146-020-01087-8

BOOK REVIEW

AI ethics – a review of three recent publications

Johann-Christian Põder[1]

Received: 12 October 2020 / Accepted: 12 October 2020 / Published online: 6 November 2020

In recent years, AI has become a hotly debated topic across different disciplines and fields of society. Rapidly advancing technological innovations, especially in areas such as machine learning (as well as increasingly widespread uses of AI-based systems), have brought about a growing awareness of the need for AI ethics, whether in politics, industry, science, or in society at large. In the following, I consider three recent publications that aim to meet these needs. Each of these books stems from a European context: one is written in English; two in German. This selection should help us to better understand the international scope of AI ethics, while fruitfully highlighting certain characteristics of each book. I begin with an overview of each publication, thereafter moving on to more general remarks and reflections.

The first publication is written by Mark Coeckelbergh, who is Professor of Philosophy of Media and Technology at the University of Vienna, Austria. His recent book *AI Ethics* (2020) was published in the Essential Knowledge Series of the MIT Press, which aims to offer accessible and expert overviews of various timely topics. Coeckelbergh aims to go beyond AI "dreams and nightmares", instead critically examining diverging assumptions about AI and humans, as well as focusing on existing AI and its current applications.

Coeckelbergh begins with a discussion of the possibility of general AI. He does not take an explicit position, and remains rather circumspect (p.66; p.141). He highlights that such debates reflect divergent understandings of the human being, as well as "deep divides" in modernity. The so-called New Romantics, argues Coeckelbergh, stress the mystery of the human being against the Enlightenment's efforts to explain it away, while humanists stress the value of the contemporary human being against transhumanist enhancement projects. The final divide—that which exists between humanism and posthumanism—is, on Coeckelbergh's account, an "interesting direction to explore" (p. 42). Inspired by the arts and humanities, posthumanism questions the ontological centrality of the human and shows a way beyond the competitive Western narrative of humans and machines. It also opens up the possibility that AI does not need to be similar to us (in either its essence, intelligence, or creativity), and can be a productive partner engaged in genuine collaboration.

Next, Coeckelbergh addresses the issue of AI's moral status, both in terms of moral agency and patiency. Here again, Coeckelbergh seeks to describe the debate rather than staking out a particular position therein. He nonetheless emphasizes a relational, socially-embedded approach to moral status, which is neither abstract nor formalizing, nor is it based upon superior, hegemonic attitudes (p.59). Thus, on Coeckelbergh's account, AI ethics forces us to reconsider our peculiarly human moral attitudes and to question our human nature and future.

Having discussed such basic issues, Coeckelbergh turns to more practical questions posed by current and near-future AI (i.e., narrow AI). He focuses on areas of ethical concern, including privacy and data protection, manipulation and exploitation, fake news, totalitarianism, safety and security, responsibility and transparency, bias, and the future of work. In addition, he discusses AI policymaking and its challenges, especially elucidating European regulations. Finally, Coeckelbergh explores the question of whether AI ethics should be anthropocentric, thereby relating AI ethics to the problem of climate change.

The second publication, *Roboterethik: Eine Einführung* 2019, [*Robot Ethics. An Introduction*], is written by Janina Loh, who works at the University of Vienna (and is a colleague of our foregoing author, Mark Coeckelbergh). Since it deals, to a great extent, with AI-based intelligent robots,

Mark Coeckelbergh (2020). AI Ethics. Cambridge, MA: MIT Press. ISBN: 9780262538190

Janina Loh (2019). Roboterethik: Eine Einführung. Frankfurt am Main: Suhrkamp Verlag. ISBN: 9783518298770

Catrin Misselhorn (2018). Grundfragen der Maschinenethik. Stuttgart: Reclam Verlag. (4th Reviewed and Revised Edition 2019). ISBN: 9783150195833

✉ Johann-Christian Põder
johann-christian.poder@uni-rostock.de

[1] Faculty of Theology, University of Rostock, Universitätsplatz 1, 18055, Rostock, Germany

Loh's work can also be read as an introduction to AI ethics. She notes that, in germanophone discourse, robot ethics has not yet established itself comparably with the anglophone world, and is often regarded with skepticism. Alongside an attempt to accomplish the above, Loh's main aim is to show the possibility of an inclusive robot ethics, which overcomes the discriminatory and hegemonic implications of traditional approaches.

Loh first focuses on what she calls the two traditional "working fields" of robot ethics, viz., on robots as moral agents and patients, lucidly presenting a spectrum of concepts from the current international discussion (not least in helpful, comparative illustrations: p.73; pp.94-5). Loh's own view is inspired by Wendell Wallach and Colin Allen's concept of "functional morality". In a third step, she discusses "inclusive approaches", which are critical of the traditionally essentialist and anthropocentric positions (pp.95–120), and often converge with critical posthumanist, feminist, and poststructuralist positions. This entails a *relational* approach that is inclusive of both humans and nonhumans vis-à-vis capabilities and attributes. Important inclusive authors for Loh are, e.g., David Gunkel and Mark Coeckelbergh.

Each of these working fields is subsequently illuminated in the light of the issue of responsibility. Loh partially defends a classical conception of individually-anchored responsibility, which should help us gain clarity and orientation in complex moral contexts. The "operational" or "functional" responsibility of robots should be complemented with human responsibility, which Loh describes as a "responsibility network". At the same time, she aims to outline a critically posthumanist and, in a deep sense, *relational* concept of responsibility, which does not focus on a 'monadic' subject and its attributes but on inclusive interaction and otherness. Her book ends with an engaged plea for inclusive and critical discourse that is not radical but open to classical, exclusive positions in robot ethics.

The third book, *Grundfragen der Maschienenethik*, 2018 [*Basic Questions in Machine Ethics*], is written by Catrin Misselhorn, who is Professor of Philosophy at the University of Göttingen, Germany. Machine ethics focuses on ethics *for* machines; on questions regarding whether machines can or *should* be capable of moral action; and how such machines should, or can, be constructed. In her instructive and well-written book, Misselhorn illuminates a theoretical foundation for machine ethics, discussing machines as moral agents and the associated implementation of morality, thereby examining three application areas for intelligent machines (viz., intelligent care robots, military robots, and autonomous vehicles).

Misselhorn is clearly skeptical about a scenario in which intelligent machines become—in the foreseeable future—full moral agents that are capable of consciousness, free will, and responsibility. She develops, however, a graduated concept of "functional moral agency" as a proper, realistic working field for machine ethics. It operates with "quasi-intentional" inner states, which are functionally comparable with human mental and moral capabilities, although less complex and greatly circumscribed (pp.86–87). Against this background, Misselhorn explores the implementation of morality as a core subject of machine ethics, discussing different top-down, bottom-up, and hybrid approaches. Using care robots as a case-in-point, she argues for a hybrid system based not only on expert opinions but on users' preferences and values.

Misselhorn's discussion of the three application areas of AI—care robots, military robots, and autonomous cars—contains fairly detailed analyses of different arguments and texts (on military robots, e.g., see pp. 155–184). Her general conclusion is that, while care robots can be ethically justified in certain contexts, military robots (as well as autonomous cars) face serious ethical objections, for questions of life and death are, in these contexts, salient (e.g., there exists a so-called "responsibility gap"). This mixed result mirrors Misselhorn's conviction that ethical issues in machine ethics cannot be resolved on a general level: they require examination in their specific application contexts.

Misselhorn's book concludes by critically prospecting about the question of "singularity" (an addendum to the 4th edition). In a critical discussion of David Chalmers' ideas (brain simulation, artificial evolution, and "the hard problem of consciousness"), she rejects the view that we will—in the foreseeable future—experience the emergence of "singularity" or "superintelligence". For Misselhorn, debates about singularity are misleading, for the biggest threat to humanity is, in fact, climate change (p. 214), although this is not pursued by Misselhorn in any detail in the context of machine ethics.

With this overview, we can now highlight some similarities and differences between these three publications. We may also point to some strengths, as well as possible weaknesses, that each text encounters. First, these publications share some important similarities. All three publications are attempting to move beyond "nightmare" scenarios and hype in relation to AI, focusing instead on existing and near-future technologies, as well as on real-life ethical issues. They are all cautious or skeptical about the possibility of general AI and the much-peddled "transhumanist science fiction" (Coeckelbergh) concerning singularity. At the same time, they favor a middle position that ascribes to intelligent machines some form of moral agency or status. Indeed, both Loh and Misselhorn develop, e.g., a view of "functional moral agency", while Coeckelbergh and (again) Loh advocate for a relational, inclusive concept of moral status which does not rely upon (objective) morally relevant properties. Furthermore, each author works through and highlights the critical insight that AI ethics is not only about technology;

it is fundamentally about *us*, viz. how we understand ourselves as human beings, as well as what kind of future we may wish for.

In spite of such similarities, all three books have their own distinctive features and advantages. Whereas Loh and Misselhorn address both the scientific community and the wider audience, Coeckelbergh aims to provide an easily-accessible overview for non-specialists. Drawing upon his extensive experience and expertise in the field of AI ethics, and in the ethics of technology more generally, he masters this task with remarkable skill, wit, and sovereignty. Overall, Coeckelbergh's approach is characterized by three distinctive and compelling, closely-intertwined features: collaboration; relationality; and vulnerability.

First, Coeckelbergh seeks to overcome a narrowly anthropocentric perspective, thereby moving towards a collaborative view of humans and AI. To this end, he is sympathetic with such approaches as posthumanism and postphenomenology, as well as the non-Western tradition, e.g., Japanese cultural attitudes towards technology. Second, his book stresses a relational, embodied, and situational approach, both to AI and ethics. This enables an inclusive openness towards AI (e.g., regarding its 'moral status') but also points to the limits of AI (e.g., its lack on 'practical wisdom'). A third, central ethical perspective concerns vulnerability. Coeckelbergh stresses that human beings are "existentially vulnerable", arguing that "AI can deny our vulnerable, bodily, earthly, and dependent existential condition" (p.196).

In addition to these features, Coeckelbergh's book is characterized by an excitingly wide thematic scope. It includes three subjects, with which Loh and Misselhorn do not grapple: AI and religion; AI and policymaking (mirroring Coeckelbergh's own professional experience); and AI and climate change. These elaborations contain valuable insights, enabling us to get into view the broader implications of AI ethics. Some subjects may have deserved, however, a lengthier treatment (e.g., non-Western approaches to technology, as briefly presented in Coeckelbergh's second chapter).

As with Coeckelbergh, Loh is critical about traditional, anthropocentric approaches, instead supporting a *relational* view of humans and intelligent robots. However, she uniquely advances a decisively critical-feminist view of robot and AI ethics. For Loh, traditional discussions of moral agency and patiency are often marked by discriminatory, exclusive views, representing patriarchal, Western, heteronormative, and white biases. Loh's posthumanist and feminist position is, however, critically open to classical, essentialist positions within robot ethics. She points out that inclusive, posthumanist positions should not be seen as a total replacement of essentialist positions, but rather as a critical extension or complement thereto. Her book can be seen as interestingly combining and balancing traditional and posthumanist approaches, while also reflecting a genuine sympathy and preference for the latter. In both Loh's and Coeckelbergh's monographs, there is a reassuring echo of Donna Haraway's *Cyborg Manifesto*.

A further distinctive feature of Loh's book is that it presents robot ethics as an ethics of responsibility. Although Coeckelbergh and Misselhorn also focus on the issue of responsibility, it does not define the conceptual outlook of their monographs. Thus, those interested in fundamental (not to mention *increasingly urgent*) questions of responsibility in relation to AI and robots (e.g., of its attribution and distribution) will find Loh's book a valuable source to this end. Nevertheless, Loh's introduction to robot ethics remains fairly theoretical, addressing in detail fundamental, 'ontological' issues. As such, Loh does not really focus on those practical ethical questions that are urgent in today's robot ethics, such as privacy and data protection, bias and discrimination, the environment, and the future of work. In contrast, such practical issues are addressed well in Coeckelbergh's book.

By contrast with Loh and Coeckelbergh, Misselhorn's lucidly written book neither reflects nor addresses feminist or posthumanist aspirations. Informed by the philosophy of mind, she carefully examines different morally-relevant mental states and human properties in relation to (intelligent) machines, without thereby attempting to cross or blur the ontological boundaries between machines and humans. Her book moreover includes two distinctive and helpful features that cannot be found in either Loh's or Coeckelbergh's introductions. First, Misselhorn presents ethical theories that are often central to machine ethics (in particular, utilitarianism, Kantianism, and virtue ethics). She thereby explains how these approaches can be fruitfully employed. Second, Misselhorn offers an in-depth discussion of some central application areas of machine ethics. Although this may contribute to Misselhorn's monograph being, in part, overly detailed, it is indeed thanks to this feature that her book convincingly combines theoretical and practical questions. At the same time, the scope of Misselhorn's book is somewhat limited, since machine ethics regards ethics *for* machines. This covers terrain such as reflecting upon machines' (possible) moral agency and decision-making abilities. This means that the issue of moral patiency—as well as many other questions—do not fit into the framework of this book. As her later added reflections about our moral attitudes towards machines reveal, this definitional focus can be hard to maintain, especially in an introductory book.

In summary, then, all three books can serve as solid and accessible introductions to AI ethics, and can profitably be read both by the scientific community and by a wider public. They offer valuable insights into a wide range of ethical issues relating to AI, thereby inviting further reflection upon how AI is shaping and changing our present and future lives. All three books fruitfully and critically complement one

another, helping us to see how AI ethics oscillates between anthropocentric and posthumanist approaches, thereby aiming to find ways to ethically articulate and conceptualize issues and developments that humanity is yet to face. To tackle these issues, we need an equal and participatory discourse (Loh), practical wisdom (Coeckelbergh), and the right moral attitudes towards those who do not belong to our species (Misselhorn). Together, these closing remarks from each of the authors offer an inspiring vision and impetus for further reflection and development of AI ethics.

Funding Open Access funding enabled and organized by Projekt DEAL.