



The Ethics of Artificial Intelligence: Review of *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI* by R. Blackman; *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges* by B.C. Stahl, D. Schroeder, and R. Rodrigues; and *AI Ethics* by M. Coeckelbergh

Ethical Machines: Your concise guide to totally unbiased, transparent, and respectful AI, Harvard Business Review Press, 2022, 224 pp., ISBN 9781647822811; *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*, Springer Nature Switzerland AG, 2023, 116 pp., ISBN 9783031170409; *AI Ethics*, The MIT Press, 2020, 248 pp., ISBN 9780262538190

Christian Goglin¹

Received: 27 April 2023 / Accepted: 12 September 2023 / Published online: 16 October 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

The ethics of artificial intelligence (AI) is a fast-growing, multidisciplinary research topic at the crossroads of engineering and social sciences. The steady increase in scientific articles and books on the topic appears to follow the curve of interest in AI, from the scientific community, the business world, regulators, and the public (Haenlein & Kaplan, 2021). Questions of responsibility, norms, and trust appear even more crucial today when we consider the emergence of new AI models with spectacular capabilities, which promise the advent of a new industrial, or even civilizational, revolution. We have summarized and cross-referenced three recent works on the ethics of AI to give the reader an overview of the range of topics currently covered by this discipline. These books have been chosen on the basis of their complementary approaches: (1) implementing an operational framework to mitigate AI ethical risks, (2) understanding the variety of socio-economic consequences of AI systems thanks to many AI ethics case studies, and (3) reflecting on the political, anthropological, and philosophical perspectives induced by the impressive progress of this technoscience.

Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI, the first book in this brief literature review, is the least theoretical of the three, with almost no references to the scientific literature. This is explained by the intended audience of the book, which presents itself as a practical handbook for policy makers, engineers, and other professionals involved in the design, development, and deployment of AI. Each chapter concludes with a short summary, listing the key elements to be retained, which again points to the practical nature of the book. Because of this objective of practicality, the author, Reid Blackman, originally a professor of philosophy at Colgate and UNC-Chapel Hill Universities, who has since become a consultant and lecturer, deliberately avoids dealing with the philosophical currents in ethics, as well as speculative digressions on future and sometimes fantasized dangers, such as the advent of a super intelligence, i.e., a system with a consciousness (Kaplan & Haenlein, 2019).

We'll say it right away, the book achieves its objective. It popularizes complex concepts and proposes the development of a clear governance framework for managing ethical risks in organizations, adopting the perspective, and sometimes the vocabulary, of the business world (dashboard, KPI, etc.).

The book uses real-life examples to justify how important it is for a for-profit organization to seriously consider the ethics of AI, particularly in terms of the critical risks involved, including legal, reputational, and ultimately, financial. The

✉ Christian Goglin
cgoglin@groupe-igs.fr

¹ ICD Business School, 12 rue Alexandre Parodi, 75010 Paris, France

author also distinguishes between AI for Good, which pursues positive societal goals, and AI not for Bad, which is of interest for commercial organizations wanting to use AI for the more prosaic purpose of productivity gains. The issue of AI not for Bad is considered to be a risk management question, aimed at identifying and mitigating risks. This vision can be compared to the stance advocated by the European Commission in its AI Act, currently under development, which aims to establish a regulatory framework to promote trustworthy AI (Artificial Intelligence Act, 2021). To achieve this, it proposes a risk-based approach in which the regulatory requirements differ depending on the risk category. Another important point in the book is the clear distinction between structure and content, with structure referring to the organizational framework and governance processes for ethical risks, and content referring to the industry-specific nature of the risks at stake. In short, content refers to the goals, what ethical risks to mitigate, while structure refers to the strategy for achieving these goals, how to mitigate these risks.

Although the author does not enter into philosophical considerations of ethics, he endeavors to clarify the concept, which is often portrayed as subjective, making it inherently unsuitable for any kind of operationalization. Even if this choice is a deliberate attempt to offer a practical, jargon-free manual, it is nevertheless regrettable that the author does not at least explain the basics of two of the three main Western ethical traditions, deontological and utilitarianism (the third is virtue Ethics), on which his approach is implicitly based. The concept of value is also presented as central and necessary to the development of an ethical framework. The author clearly explains what an operable ethical value is, enabling us to go further than the abstract value statements sometimes found in standard AI ethics statements.

Having laid these foundations, the author then devotes a chapter to each of the three major risks commonly discussed in connection with Machine Learning, the “big three”: bias, explainability, and privacy (Davenport et al., 2020; De Bruyn et al., 2020). In each case, the explanations and examples are clear and understandable for the public and demonstrate a thorough knowledge of these technical subjects. Along the way, the author gradually builds an organizational and governance framework for risk identification and mitigation. In addition to the acculturation measures needed to get employees on board, the author recommends “ethics by design” and the establishment of a multidisciplinary AI ethics committee comprising engineers, lawyers, senior managers, and ethicists. Another interesting recommendation is to keep a register of decisions taken in the face of previously encountered ethical dilemmas, to create a body of “case law” specific to the company.

Going one step further in developing a practical framework for managing the ethical risks of AI, the book proposes

a method for identifying key ethical values, linked to the fundamental risks of the activity concerned. These values are intended to serve as a compass for management. In this approach, the ethical values are justified (the why), the actions for mitigating the related risks are made explicit (the what we do), and the concrete applications of these actions are presented (the how).

Finally, the author concludes by explaining that, with the exception of the specialized chapters, dealing with AI-related risks such as big three, and replacing “AI ethics” with “ethics”, the remaining chapters are, in themselves, a general method for managing ethical risks of all kinds.

The second book in our review takes a very different position, not at all comparable with *Ethical Machines* and not aimed at the same audience. Well documented, with numerous references to the scientific literature, *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges* is co-authored by B.C. Stahl, D. Schroeder, and R. Rodrigues. The first author is Professor of Critical Research in Technology at the School of Computer Science of the University of Nottingham. This compact book analyzes the different facets of AI ethics through a series of real-life case studies, notably from the EU-funded SHERPA research project (2018–2021).¹

The book is systematically structured. Each chapter focuses on a specific theme and starts by presenting several illustrative examples, followed by a broader discussion that leads to solutions for mitigating the ethical risks in question. These solutions are drawn from the scientific literature and the major emerging regulations on AI. They are based either on tools or on a reasoned system of decision-making supported by ethics. Unlike *Ethical Machines*, which deliberately avoids referring to the major Western ethical traditions (Aristotle’s virtue ethics, Kant’s deontology, and Bentham’s consequentialism), *Ethics of Artificial Intelligence* claims them as a basis for reasoning, with a pluralistic approach and a slight emphasis on deontological ethics in several chapters.

The illustrative examples drawn from real-life events are one of the main assets of this book.

First of all, they make the sometimes abstract ethical risks concrete. The example of Nijeer Parks, an individual arrested for a crime he did not commit based on a matching error, is striking. This type of story, with dramatic consequences, is likely to raise awareness of the risks associated with AI, and undoubtedly makes more of an impact than a theoretical statement noting that the model’s discriminatory nature can be attributed to the under-representation of a particular category in its training dataset.

¹ This research received funding from the European Union’s Horizon 2020 Framework Programme for Research and Innovation under Grant Agreements No. 786641 (SHERPA).

These cases sometimes allow us to extend our reflections on AI. For example, with respect to privacy, the authors relate the case of a genetic research project carried out in Saudi Arabia, where it was revealed that the subjects in the sample were not aware of the risks associated with the disclosure of their health data. In this case, the nature of the data (genetic), and the fact that these data revealed information not only about the individual but also more broadly about the individual's entire family (which shares part of the same genome), highlights an unexpected risk—the potential impact on a family when one of its members participates in a study.

The multiplicity of ethical issues raised in the cases presented is surprising and offers a broad and instructive panorama of ethical risks. We note the well-known cases of discrimination by automated hiring systems or predictive policing, the problems of violation of privacy by authoritarian regimes using data for social scoring purposes, the concept of surveillance capitalism related to tech giants appropriating massive data to monetize them, and the cases of electoral manipulation enabled by AI. Other themes are less expected, notably the right to security in cases where home automation systems are hacked, including the control of heating systems; the falsification of medical diagnoses by AI that generated false radio images; and the thorny ethical issues associated with medical robots and, even more so, with sex robots (Ma et al., 2022).

The authors propose concrete solutions, with several references to the work of the European Commission High-Level Expert Group and its self-assessment tool (HLEG, 2020). This risk mapping approach echoes the one described in *Ethical Machines*, which uses it as a starting point for the development of an operational ethical framework. Another solution shared by both books is the adoption of ethics by design, an approach also inspired by the HLEG. Ethics by design implies the consideration of an explicit value system, used in the design and development phases.

The analysis of these multiple cases also shows that many of the ethical problems highlighted with respect to AI are not new, but in fact pre-existed its advent, discrimination being a good example. This is not to say that AI is an ethically neutral technology. It brings new and singular risks, such as the opacity of deep learning models or the need to collect ever-increasing amounts of training data, which create an incentive to disregard moral and legal requirements relating to privacy. It is worth noting that this position—AI is not ethically neutral—is another point of agreement with the previous book.

With respect to the risks inherent in Machine Learning—in other words, the big three—*Ethics of Artificial Intelligence* links explainability and discriminatory bias, which *Ethical Machines* fails to do. Obtaining an explanation for an automated decision is likely to provide information on

whether it is discriminatory. Accordingly, the GDPR (recital 71) provides for the “right to explanation” for any decision that has a significant impact on citizens' lives, so that they can appeal if discrimination is found.

The authors also note that it is often not so much the AI itself that poses an ethical problem but the purpose of its use as well as its integration in a technical system itself inserted in a social system. This systemic link complicates the ethical analysis because the consequences at stake must be analyzed in light of this socio-technical context, which includes not only laws and regulations, but also social and moral preferences.

Finally, one of the concluding points put forward by the book appears particularly important: balancing of the values at stake in the ethical dilemmas stemming from the use of AI. Several cases in the book are concerned by this point, predictive justice for example, where there may be a tension between the will to reduce crime through better prediction (safety value) and the risk of discriminating against disadvantaged people (equity value), as revealed by Angwin et al. (2016). These ethical dilemmas raise the question of how to balance competing values. Such issues are a matter of societal choice and require political resolution.

The third book, *AI Ethics*, again offers a different and complementary position, providing a broad overview of the ethical questions raised by the current and future uses of AI. Numerous reflections extensively address more global, and metaphysical, questions such as the advent of a super intelligence, transhumanism, and the future of humanity and the planet. Its author, Mark Coeckelbergh, is Professor of Philosophy of Media and Technology at the University of Vienna.

The book begins by examining ancient myths (Pygmalion, Golem, Prometheus, etc.) and more recent fictions (Frankenstein; 2001: A Space Odyssey; Terminator; Ex Machina, etc.) that shape our imaginations and the imaginations of Silicon Valley entrepreneurs. These narratives form the bedrock of our hopes, fantasies, and fears regarding the advent of a super intelligence, and it is useful to revisit them. Going further in these comparisons, the author links several AI-related themes, such as transhumanism or singularity, to the great Western religions by showing that they share certain beliefs such as transcendence and immortality. Various fascinating metaphysical and philosophical questions are also explored, including the fundamental difference between man and machine and the possibility of strong AI. The author then goes on to outline the different views on machines as moral agents and how they should be treated. Put succinctly, should AI technologies be viewed in the same category as a toaster, or should they one day be granted rights? These controversies divide the scientific community, mixing speculation, scientific knowledge, and moral philosophy, but they are not without merit in our opinion, given the rapid

and unpredictable technological progress we are currently witnessing.

The book then leaves the register of philosophical speculation and, after a historical review of AI, focuses on the ethical issues raised by the impacts of current applications of this technology. The view given on the ethics of AI has analogies with that of *Ethics of Artificial Intelligence*, insofar as the applications of AI must be thought of in their technical, social, and historical context. Naturally, the big three are explained and illustrated. The literature review of ethical risks is exhaustive and its analysis is dense, articulating many dimensions: legal (with the question of responsibility), social (with the risks of discrimination), societal (with the risks to democracy), anthropological (with the risks of human dependence and vulnerability), economic (with the questions of employment, business secrecy, and the link between explicability and trust), and political (what justice for what equity, the environmental question, etc.). Other stimulating reflections concern possible systemic evolutions or the meaning of life in a future society in which work has practically disappeared.

Another strength of the book is that it addresses issues relating to the regulation of AI by exposing the complexities involved in preparing such regulation, including: the need for justification on the basis of principles and values; the paramount importance of the definition of AI; the different temporalities of the law and its proactive or reactive character; the right balance between the ethical question and the economic and geopolitical stakes; the treatment of biases relative to different visions of justice, etc. The author also shows that a consensus is emerging from the profusion of norms and regulations being proposed around the world, with a focus on the importance of addressing the existing problems posed by AI rather than more speculative issues currently more akin to science fiction. Another notable point is that the themes addressed converge (the big three, etc.) as do the solutions provided (ethics by design, etc.). For the author, these efforts should be made more inclusive by using a bottom-up approach, taking into account the opinions and interests of the various stakeholders, through public debates, focus groups, etc. Moreover, their implementation comes up against multiple challenges: the GAFAMs' concentration of power, the difficulty of translating virtuous but inherently imprecise principles into operational standards, and the near impossibility of stopping the development of innovations that run counter to moral principles.

In response to these challenges, the author makes a strong plea for more education, multidisciplinary, democratic debate, and openness to philosophies other than Western philosophy. The key point, he argues, is to see ethics as a positive reflection, necessary for a good life, rather than seeing it simply as a coercive normative force. Finally, after discussing the excesses of an overly anthropocentric vision and

the need to include the planet in the debate, the author concludes this exciting journey by noting that humanity needs to act wisely when facing the considerable challenge of AI.

To conclude, although these three books address the same subject, the ethics of AI, their respective objectives are different. *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI* is a practical handbook on ethical risk management for professionals. It is easy to read and not particularly conceptual or theoretical; however, it is not very well documented. That being said, the pedagogical quality of the book and its practical recommendations seem useful, which is why we recommend it to professionals and managers involved in the design, development, or use of AI. In addition, the risk management framework that it presents could be used to conduct tutorials for computer science or management students. It would be even more interesting to mix this framework with the risk mitigation solutions of the second book, in particular the HLEG's self-assessment tool for trustworthy AI.

The second book, *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*, will be of more interest to researchers, students, or individuals wishing to benefit from a broader perspective on the current ethical risks of AI in the context of its ever-expanding real-life applications. This book does not confine itself to the perspective of for-profit organizations, but widens the focus from the individual to society and the regulator.

Given its structure as a bank of real-life case studies, it is also very useful as a resource for objectifying and illustrating the different ethical risks linked to the uses of AI, thanks to its striking examples. It also offers perspectives for resolving these issues, drawn from both the scientific and regulatory communities.

Finally, the third book, *AI Ethics*, is perhaps the most thought-provoking in its openness and its in-depth analysis of the current and potential consequences of AI, which is an approach we can connect to the call by Hermann (2022) to base the analysis of the AI usages on utilitarian perspective weighing benefits and costs across stakeholders. The richness of the book, addressing myths, history, societal aspects, and speculative and metaphysical perspectives, complements the second book's lessons from current applications. It is also important to consider the future risks of AI today to steer the future in the right direction (Letheren et al., 2020). Indeed, technological progress in this field, which attracts significant levels of funding, appears to have accelerated over the last two decades. Reflections based solely on the current situation, with no attempt to integrate reasonable future scenarios, carry the risk of producing rapidly obsolete recommendations that are incapable of dealing with new situations. Moreover, as we mentioned in the introduction, it is apparent that the emergence of AIs approaching the concept of artificial general intelligence (Kaplan &

Haenlein, 2019), at least in appearance, call into question the boundaries of what is possible. Conversational agents based on large language models and recent deep neural network architectures (transformers) appear to be on the verge of disrupting the way we organize the world of work and value chains and raise dizzying anthropological, societal, and political questions.

In this march of history toward an increasingly technological world, we undoubtedly need ethics more than ever. Each of these books contributes, in its own way, to meeting this need.

Funding No funds, grants, or other support was received.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The author has no financial or proprietary interests in any material discussed in this article nor non-financial interests.

Ethical Approval Not applicable.

Informed Consent Not applicable.

Research involving Human Participants and/or Animals Not applicable.

References

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. Retrieved from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1), 24–42. <https://doi.org/10.1007/s11747-019-00696-0>
- De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J.K.-U., & von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, 51, 91–105. <https://doi.org/10.1016/j.intmar.2020.04.007>
- Haenlein, M., & Kaplan, A. (2021). Artificial intelligence and robotics: Shaking up the business world and society at large. *Journal of Business Research*, 124, 405–407. <https://doi.org/10.1016/j.jbusres.2020.10.042>
- Hermann, E. (2022). Leveraging artificial intelligence in marketing for social good—an ethical perspective. *Journal of Business Ethics*, 179(1), 43–61. <https://doi.org/10.1007/s10551-021-04843-y>
- High-level expert group on artificial intelligence. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment* (p. 34). European Commission.
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- Letheren, K., Russell-Bennett, R., & Whittaker, L. (2020). Black, white or grey magic? Our future with artificial intelligence. *Journal of Marketing Management*, 36(3/4), 216–232. <https://doi.org/10.1080/0267257X.2019.1706306>
- Ma, J., Tojib, D., & Tsarenko, Y. (2022). Sex robots: Are we ready for them? An exploration of the psychological mechanisms underlying people's receptiveness of sex robots. *Journal of Business Ethics*, 178(4), 1091–1107. <https://doi.org/10.1007/s10551-022-05059-4>
- 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*, 108 (2021) (testimony of Commission European). Retrieved from <https://artificialintelligenceact.eu/the-act/>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reproduced with permission of copyright owner. Further reproduction prohibited without permission.