Pro-Fil 25 (2) 2024: 69–71
doi: https://doi.org/10.5817/pf24-2-39786

SHANNON VALLOR
THE AI MIRROR: HOW TO RECLAIM OUR HUMANITY IN AN AGE OF MACHINE THINKING

Oxford University Press, 2024, 257 p.

JAKUB PELOUŠEK

Department of Philosophy, Faculty of Arts, Masaryk University, Brno, Czech Republic,
pelousek@mail.muni.cz

BOOK REVIEW

Shannon Vallor is an American philosopher of technology. She holds the Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence at the Edinburgh Futures Institute (EFI) at the University of Edinburgh and works there at the Department of Philosophy. Besides this, she is the Director of the Centre for Technomoral Futures within EFI and co-director of the BRAID (Bridging Responsible AI Divides) program, funded by the Arts and Humanities Research Council. Vallor is also a standing member of the One Hundred Year Study of Artificial Intelligence (AI100) and serves on the Oversight Board of the Ada Lovelace Institute. On top of that, she has received several awards, including the 2015 World Technology Award in Ethics and the 2022 Covey Award from the International Association of Computing and Philosophy.

Her research examines how emerging technologies, particularly AI, robotics, and data science, shape human moral character, habits, and practices. She frequently advises policymakers and industry leaders on ethical approaches to AI design and application. In addition to publishing numerous articles and educational resources on data, robotics, and AI ethics, she is the author of *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford University Press, 2016) and *The AI Mirror: Reclaiming Our Humanity in an Age of Machine Thinking* (Oxford University Press, 2024), the focus of this review.

The main theme of this book revolves around the metaphor of the mirror. AI technologies mirror human intelligence – our thinking, judgments, biases, desires, needs, expectations, imaginings, and more. Vallor highlights the case of Blake Lemoine, a Google engineer who mistakenly interpreted the output of the LaMDA language model as evidence of a sentient and self-aware being. However, Vallor argues that this is merely an illusion, as contemporary AI systems ‘are constructed as immense mirrors of human intelligence’(2).

This reflected image, however, can be dangerous. Vallor illustrates this with the story of Narcissus and Echo from Book III of Ovid’s *The Metamorphoses*. Narcissus, known for his beauty and pride, became captivated by his own reflection in a pool, falling in love with the illusion of his image. Similarly, Vallor emphasizes that we are fascinated by AI because it represents a less fragile, seemingly more perfect version of our own abilities.

In the first chapter, Vallor emphasizes that AI systems and advanced computing technologies are becoming magnifying mirrors of humanity. What we give to them is reflected back with amplified power. One could argue that all the errors in AI systems are merely human errors, as humanity serves as the blueprint for these systems. They are nothing more than amalgamated projections of our past, reflected in countless variations. This illusion leaves us stuck, unable to progress, in the face of contemporary crises such as climate change or global political instability.

However, in the second chapter Vallor asserts that all is not lost. While AI systems lack their own morality, they are not neutral – they are based on the moral beliefs of their creators - humans, but they represent only a small minority of humanity. This makes it essential to raise ethical questions about AI and adopt a responsible approach to its development and use.

In the third chapter, she also references several science fiction authors, including Samuel Butler and Karel Čapek, as examples of how literature addressed similar questions decades before they became pressing concerns in the real world. Vallor further highlights that AI systems, such as language or predictive models, are not technologies to be feared. They operate based on human logic and function as extensions of our own abilities, representing weak or narrow AI rather than any form of general or superintelligent AI.

Vallor adopts a neutral stance in the polarized debate surrounding AI. While she acknowledges concerns about the rise of AI systems, she also views them as opportunities to better understand ourselves and shape a brighter future. Despite this balanced approach, she remains critical of longtermists – those advocating for the allocation of charitable resources to AI safety research aimed at preventing the emergence of artificial general intelligence (AGI), which they perceive as a potential existential threat. Vallor argues that longtermists misunderstand the nature of AGI, as well as the stakes and potential outcomes, leading to an incomplete perspective.

The fourth chapter discusses how we should use AI mirrors as tools to prevent the repetition of past injustices and to create a better future. For example, AI can help address previously unrecognized injustices, improve the allocation of care and resources, and make scientific knowledge more accessible to humanity. She critiques the longtermist doctrine, noting that rather than enhancing AI safety, it exacerbates imbalances in AI transparency and accessibility. The real danger, she argues, lies in the narrow group of individuals developing these systems—people who represent only a small subset of society. Because AI mirrors its creators, it risks reflecting their attitudes and biases rather than providing a universal representation of humanity.

Contrary to the portrayal of AGIs in some science fiction, such as Ava from *Ex Machina* or the replicants in *Blade Runner*, which depict AGIs as cold, calculating machines, Vallor imagines AGIs as curious and playful entities. This perspective, developed in the fifth chapter, is rooted more in human characteristics than in mechanistic ones and presents a more optimistic and pleasant vision of potential AGI. Vallor's view offers a refreshing alternative to seeing AI as either humanity's ultimate saviour or destroyer, offering a more comfortable and hopeful attitude toward future AI developments. Ultimately, she suggests that instead of fearing AI, we should use it as a means to learn about ourselves and make ourselves better.

The sixth chapter corresponds with Vallor's earlier work, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*, and further develops her project of virtue ethics. In this context, she emphasizes the need to 'develop *technomoral* wisdom and expertise' (166).

Technomoral virtues are the specific traits and qualities people need to navigate wisely and thrive in the uncertainty and complexity of a fast-evolving technosocial environment. Building on this foundation, she advocates for incorporating new ethical frameworks into AI mirrors to propel society forward, rather than perpetuating the current sociotechnical focus on efficiency and production

In the last chapter, she summarizes the previous parts of the book in the context of her own perspective on the morality of technology. She proposes that we should oversee the developers of AI systems for safety, ethical, and political reasons. The problem, as she states in correspondence with an essay *Man the Technician* (José Ortega y Gasset, 1961) is that technologies are not neutral; they are shaped by those who create them. AI systems should be designed to serve a good purpose, such as summarizing scientific ideas, but instead, they are prone to generating internet garbage and similar content. However, Vallor emphasizes that it is not too late to act, and we should focus on developing AI systems that are more beneficial for humanity.

The AI Mirror is aimed at readers who may not be deeply immersed in AI research or its intersections with other disciplines. Shannon Vallor examines our past, contemporary, and near-future interactions with advanced information technologies, particularly AI systems, offering a measured perspective on the challenges of coexisting with AI. Despite the complexity of the topic, the book is written in an accessible style, making it suitable for beginners and readers from diverse fields.

ICT experts and researchers may find this book a valuable critical reflection on their work. Vallor approaches information technologies from a human-centred perspective that transcends the insular focus of the ICT field. She reminds readers that ‘we’ do not represent humanity as a whole, but rather a specific subset of people. Given the immense influence of AI systems, Vallor argues that it is essential to include reflections and critiques from those outside this narrow group.

Ultimately, Vallor asserts that AI systems alone cannot deliver a better future unless we change how we behave within our technosociety. Achieving this requires rebuilding and refining our moral attitudes rather than clinging to the current sociotechnical paradigm.