


Book Review

Shannon Vallor, *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*

Tomislav Furlanis

University of Ljubljana, Slovenia

tomislav.furlanis@teof.uni-lj.si

 © 2024 Tomislav Furlanis

In the famous scene of the movie *Matrix* (Wachowski and Wachowski 1999), Agent Smith climactically holds captive the severely beaten-up protagonist Neo against the tracks in a powerful clinch as the train approaches. With confidence, he proclaims: ‘Do you hear that, Mr. Anderson? That is the sound of inevitability. It is the sound of your death. Good-bye, Mr. Anderson.’ The protagonist, Neo, looking at the approaching train, responds with a quiet but resolute ‘My name ... is Neo,’ to suddenly and forcefully release himself and jump out from the oncoming train and his demise. This scene poignantly embodies the motif of the movie, which is humanity’s struggle to break free, to wake up, from the forceful surrender of their collective free will, imposed upon them by deterministic machines, made possible through an elaborately woven illusion of the matrix.

The most recent book by Shannon Vallor (2024), the famed philosopher of technology and virtue ethics, *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*, reflects a similar existential orientation. She wants global humanity to wake up from what Langdon Winner called ‘technological somnambulism,’ where we ‘willingly sleepwalk through the process of reconstituting the conditions of human existence’ (Vallor 2024, 216, quoting Winner 1977). She wants us to understand the existential dangers of following the dominant techno-deterministic narratives, predominantly propagated by big tech companies, that inadvertently diminish the essence of humanity by constraining our collective self-authorship capabilities.

There are two fundamental aspects with which Vallor engages in her book. First, she wants to show that the real existential risk of utilizing, deploying, and developing contemporary, generative AI systems is not one

of mass extinction (as in the death of the human species). Instead, the existential threat lies in the gradual loss of what makes us human. Namely, it is in the diminishing of self-determination, the gradual surrender of the will to live and to shape our civilization in creative, beneficial ways, the entrenchment of social bias, the erosion of social and moral confidence, and the denigration of the mutual care we manifest towards each other as fellow human beings. Second, she wants to highlight that the dominant AI narratives, fuelling the rapid development and investment of generative AI systems, promote these systems as inevitable steps towards the achievement of the fabled artificial general intelligence (AGI) but in all actuality only serve to reinforce the social, political, and economic power of the companies designing, deploying and selling these products to the global market. Thus, the second significant concern is the potential for social and political influence by tech giants like OpenAI and Google. These companies, with their powerful AI products and the narratives surrounding them, possess the capacity to exert considerable influence over global society.

For Vallor, when combined, these points constitute an existentially dangerous illusion, a spectre, that utterly misdirects the proper existential efforts humanity should invest in its relation to AI technologies. Here then, instead of taking human destiny into our own hands, and working it out, in trying to produce the best possible future we can envision – through the use and development of our moral and political skills – we are releasing our collective freedom and will to a technological spectre that is not only suppressing human agency into a position of existential surrender but is, in all actuality, only a proxy for the manifestation of the economic and political power of big tech companies. In Vallor's (2024, 200) words, these systems are:

designed to ensnare our attention, stoke our anger, fear, and division, and prevent us from trusting ourselves and one another to be anything more than their handmaidens. Which just means being handmaidens to the humans who build and profit from them.

Or even more clearly (Vallor 2024, 63):

Today's data-hungry tools are being built by powerful corporations to feast like insatiable parasites on our own words, images, and thoughts, strip away their humane roots in lived experience, and feed them back to us as hollow replacements for our minds.

Crucially, Vallor doesn't merely use strong language in a general sense.

She explicitly identifies those she believes are driving this harmful trend: Elon Musk, Sam Altman, Geoffrey Hinton, and the Future of Humanity Institute. These individuals and organizations are actively promoting techno-deterministic narratives while simultaneously profiting directly from the very technologies these narratives endorse and market. Similarly, she connects her arguments with the work, and social presence, of other (in)famous generative AI critics such as Gary Marcus (2024) or Timnit Gebru (Bender et al. 2021), thus positioning her message directly in the ongoing and dynamic human-AI real-life context, which, by the day, is becoming increasingly politicized in front of our eyes. Thus, the overall style of the book leans more towards a performative, not an argumentative approach. As such, the book serves better as an alarm bell than a magnifying glass.

Accordingly, interested readers, especially those equipped with the technical know-how, may be enticed by Vallor's descriptions, comparisons, and explications – but may not be impressed, as the argumentation, more often than not, rests at a perfunctory level. To exemplify, and more philosophically, the author uses the perspective of embodiment and phenomenology to point out how 'minds almost certainly come into existence through the body and its physical operations.' But the argument is left at that, without explicating why that is so. Similarly, when mentioning how, 'contrary to the suggestions of philosophers like David Chalmers, human wellbeing is inseparable from the flesh of the world' (Vallor 2024, 207), the statement does not continue to say 'why' that is so, nor are we provided with Chalmers' argument beyond this short mention. Similarly, when pointing out how the current state of AI development can be improved with the advancement of the 'responsible AI' approach, not much is added besides an introductory mention of this field of research. Or when, in portraying what specific, real-life, problems the current generative AI could be used to improve, or even solve, not much is added beyond the point that these systems could be used to narrow 'the gaps between what we need and what others have the resources or skills to give us' (p. 213).

To move the reader towards relevant action, Vallor places her ideas in seven distinct thematic chapters. The first two chapters aim to dismantle the idea of contemporary AIs as 'minds,' or 'intelligence' and thus defeat the narrative that machines can substitute humans. For, these systems are – in Vallor's description – nothing more than a "mirror", and a really bad one at that. These tools do not think or understand or reason in any mean-

ingful sense of the word, rather, they reflect the patterns found in our human thinking, understanding, and reasoning. Consequently, due to their lack of a commonsense world model, these machines are very brittle and unreliable. As a result, they generate harmful fictional accounts of real people and events, outright confabulate nonsense, and often reflect existing societal or historical prejudices.

However, while this is their reality, the systems are at the same time being promoted as efficient human performance optimizers, as systems deterministically leading us to the holy grail of a ‘superhuman AI,’ of an AGI, where ‘super’ connotes being ‘better than humans at calculation, prediction, modeling, production, and problem-solving’ (Vallor 2024, 86). For Vallor, these narratives are missing the mark, since they are promoting what Langdon Winner in 1977 called ‘reverse adaptation’ (Vallor 2024, 88, quoting Winner 1977). Here, instead of making machines adapt to humanity by becoming more humanized, humans are forced to adapt to machines by becoming more mechanized – similar to the fate of humans in the Matrix fictional universe. For Vallor, this is a great error in thinking, because it conflates human existence to mathematically describable and economically traceable performance, and it presents an obstacle to the achievement of optimal human performance. Moreover, by quoting Brynjolfsson’s concept of the ‘Turing Trap,’ Vallor argues that – even if we wanted to produce the highest performance-producing machines – we should know that machines become capable of achieving the highest performance not when they substitute human agency but instead when they augment or enhance it.

Here, however, besides the remarks, the discussion does not steer much into the foray of human augmentation or human-AI cooperation. Personally, and highlighting my own research bias (one of symbiotic cooperation), I found this lack dissatisfying as the AI field in its broadest terms has witnessed a palpable shift from ‘automation’ towards ‘augmentation’ in the past decade of both research, use, and marketing of AI systems (Jarrahi 2018; Hassani et al. 2020; Rajpurkar et al. 2022; Tankelevitch et al. 2024). As such, the point of human-LLM cooperation should have been covered more, especially when we have in mind that there are authors who are already extensively investigating the theme of human-AI cooperation (Mollick 2024).

To exemplify, Vallor points out that the moral victories humanity has achieved in the domain of civic life have been accomplished through arduous and constant engagement in the public sphere. But AI generators

are capable of diminishing ‘the value of human participation in moral and political thought [by taking] the hard work of thinking off our shaky human hands’ (Vallor 2024, 130). For it is precisely in the face of imperfection that the humans can engage the importance of the ‘meaningful moral reflection, moral appeals, moral responsibility, and imagination. What are we without these?’ (p. 129).

To combat these negative trends, we require institutions that are willing to ‘deploy planetary-scale interventions and systemic reforms of unsustainable practices, where technical and moral excellence need to come together, to synergistically engage the ongoing concerns’ (Vallor 2024, 167). AI developers cannot be omitted from the obligation of responsibility. As she concludes in the closing chapter: ‘The most plausible existential danger is not a genocidal machine oppressor but the annihilation of human moral confidence’ (p. 196). For ringing a clear alarm bell to this danger, then, this work deserves high praise and recommendation.

Acknowledgment

This publication is a part of the research programme The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights, financed by the University of Ljubljana, and research project Epistemic Identity and Epistemic Virtue: Human Mind and Artificial Intelligence, supported by the John Templeton Foundation and the Ian Ramsey Centre for Science and Religion (University of Oxford).

References

- Bender Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. ‘On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?’ In *FaccT ’21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. New York: Association for Computing Machinery, 2021.
- Hassani, Hossein, Enrico S. Silva, Sebastian Unger, Mohsen Taj Mazinani, and Stephen Mac Feely. 2020. ‘Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future?’ *AI* 1 (2): 143–155.
- Jarrahi, Mohammad H. 2018. ‘Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making.’ *Business Horizons* 61 (4): 577–586.
- Marcus, Gary F. 2024. *Taming Silicon Valley: How We Can Ensure That AI Works for Us*. Cambridge, MA: MIT Press.
- Mollick, Ethan. 2024. *Co-Intelligence*. London: Random House.
- Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J. Topol. ‘AI in Health and Medicine.’ *Nature Medicine* 28 (1): 31–38.
- Tankelevitch, Lev, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. ‘The Metacognitive Demands

- and Opportunities of Generative AI.' In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24.
- Vallor, Shannon. 2024. *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford: Oxford University Press.
- Wachowski, Lana, and Lilly Wachowski, directors. 1999. *Matrix*. Warner Bros., 2 hrs., 16 min.
- Winner, Langdon. 1977. *Autonomous technology: Technics-out-of-control as a theme in political thought*. Cambridge, MA: MIT Press.