Review of *The AI Mirror* by Shannon Vallor. Author: Daniel Story. Review forthcoming in *The Journal of Moral Philosophy*. September 21, 2024.

Shannon Vallor, *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*, (New York: Oxford University Press, 2024), 257 pages. ISBN: 978-0-19-775906-6. Hardcover: $ 29.99.

Few areas of human life remain untouched by artificial intelligence. AI influences how we work, play, entertain, communicate, learn, think, create, govern, fight, and love. Some of these influences are invisible. Many are profound. All are new. A tremendous amount of ink is being spilled about this topic. This is appropriate. Radical, abrupt change demands intensive discussion. And the demand is even more urgent because we currently face unprecedented challenges and an uncertain future.

Shannon Vallor's *The AI Mirror* is another high-altitude contribution to the discussion about the threat and promise of AI. Vallor is concerned specifically with machine learning systems that are trained on data about people, such as algorithms that assist in parole or hiring decisions, content curation algorithms, and large language models. The book's central idea is that this type of AI is like a mirror that mindlessly reflects an image of those limited portions of human life that are represented in training data. It is thus a fundamentally unthinking, conservative, and circumscribed force. Mirrors are useful, but only if they are used to the right extent, at the right time, with the right motive, and in the right way. If we use AI to determine our self-understanding and future–as it seems to Vallor that we are tempted to do–then we risk ossifying and reproducing what has come before. This would be catastrophic because we face grave global risks, such as climate change and resource depletion, that call for creative moral improvisation and self-transformation.

The AI Mirror is a natural extension of Vallor's previous monograph, *Technology and the Virtues* (Vallor 2016), in which Vallor argued that we need to cultivate new "technomoral virtues" that will enable us to flourish in our technologized and rapidly changing environment. Vallor retains this virtue-theoretic perspective and premise in her new book. But *The AI Mirror* is shorter, less technical, and more focused than the previous work, targeting recent developments in generative AI and ethics.

It is stimulating to read a commentator who acknowledges the monumental potential of AI while avoiding the standard doomer/optimist antipodes. Vallor proposes that AI takeover worries, like those articulated by Nick Bostrom (Bostrom 2014) and Toby Ord (Ord 2020), are based in colonial anxieties about power and domination. There is no reason to believe that artificial general intelligence would want to dominate us. In any case, we are not on a direct path to AGI, because unthinking mirrors cannot take us there. For this reason, Vallor is also implicitly critical of optimistic singularitarians such as Ray Kurzweil (Kurzweil 2024), who claim to see technological transcendence just on the horizon. Vallor's perspective on these antipodes leads her to argue that ethical programs like longtermism (Greaves and MacAskill 2021) that recommend we divert major charitable resources to AI research in order to either mitigate the takeover threat or usher in apotheosis only function to consolidate power and exacerbate inequality.

Sometimes critics who take issue with particular uses of a technology frame their critique as a condemnation of the technology itself. Vallor does not make this mistake. She allows that AI mirrors have uses that are conducive to technomoral flourishing. For example, they can reveal injustices that we would not have otherwise noticed, support care by connecting needy people to those with the ability to help, or disseminate scientific knowledge to the public. Additionally, Vallor makes some inspiring suggestions about how we might reimagine the guiding principles of AI design, favoring the underappreciated ideals of restraint, restoration, and care. One of the most inspiring parts of the book is the picture of AGI that Vallor offers as a counterbalance to the coldly calculating HALs (*2001: A Space Odyssey*) and Avas (*Ex Machina*) that dominate our cultural vision. Vallor's picture is of an intelligence that is silly, curious, and playful. It is reticent and supportive, something closer to Samantha (*Her*). There is no particular reason to think an AGI would not or could not have these qualities. If this picture were more prominent in our vision, if it were hoped for and aimed at, it might alter how we think about, build, and use artificial intelligence. It might even alter how we think about human intelligence and the kind of people we should try to be.

While *The AI Mirror* is by no means a polemic, the book feels heavily tilted towards the critical, leaving something to be desired with respect to ideas about the positive potential of AI. If Vallor's mirror analogy is apt, then AI fits into a rich category of cultural devices that function to reflect images of ourselves back to us, which includes rituals, institutions, oral traditions, historical methods, and art forms. Each device fashions images in distinctive ways, and these ways contribute to each device's distinctive value. For example, historical fiction and memoirs reflect images that include the interiority of some of the people depicted. This feature is part of what gives these devices their special powers and distinguishes them from, say, a chronicle, which fashions images of events but not of thoughts and feelings. We pursue various devices for gazing at our past and ourselves because, as the memoirist Mary Karr puts it, "everybody has a past, and every past spawns fierce and fiery emotions about what it means. Nobody can be autonomous in making choices today unless she grasps how she's being internally yanked around by stuff that came before" (Karr 2015 p. xxiii). This applies both individually and collectively.

It seems likely that Vallor would agree with Karr's statements. These statements suggest that AI may be distinctively valuable in the universal task of self-understanding and autofabrication (to use Vallor's preferred term). Yet Vallor does not spend as much time as one might like exploring AI's distinctive value as a reflective device. There is much to explore, though (see, for example, Haselager 2024), and exploration is necessary if we are to properly configure our relationship to AI and develop new virtues appropriate to our technologized environment.

One place where Vallor does not deviate from standard discursive assumptions is in her dim estimation of individuals' capacities to naturally avoid the most obvious misuses of AI. Developments in AI are sometimes dazzling. For example, recent advances in natural language processing have generated lots of excitement and confusion, which is understandable given the evolved human tendency to interpret comprehensible language as conveying meaning and intent (Bender et al. 2021). Like many commentators, Vallor believes this confusion is dangerous. It can lead us to think that we are talking to an intelligent agent when actually we are only gazing into a mirror. This assessment of danger depends on the assumption that people will fail to naturally correct their own confusions as the technology becomes more familiar. Fortunately, the

history of technology seems to be a cause for optimism because it suggests that fundamental confusions of this kind are often transitory. People initially misunderstood the evidential value of photography, for example. Arthur Conan Doyle was duped by the Cottingley Fairy photographs (Doyle 1921). But now we understand the limitations of photographic evidence and have adjusted our usage accordingly. It is hard to believe that we will not go through a period of similar adjustments in relation to large language models and many of the other AI technologies that Vallor is interested in, even without the radical re-envisioning she advocates. Humans have a sophisticated capacity for navigating such complexities, and we are rarely duped by our creations forever.

Ultimately, Vallor presents a valuable perspective on AI that is worth taking seriously. Moreover, Vallor's warnings about how technology can constrain self-transformation are worth attending to, even if one does not accept the premise that we urgently need to develop new technomoral virtues. As Vallor notes, AI cannot by itself lead us into a better future. It cannot by itself deliver us flourishment. But we might hope that it can be made to reflect our ugliness and our beauty in new, interesting, and virtuous ways.

Daniel Story
Department of Philosophy
California Polytechnic State University
San Luis Obispo, California, USA
dstory@calpoly.edu

ACKNOWLEDGEMENT

Thanks to Jacob Sparks for comments on a draft of this review.

REFERENCES

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 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Doyle, Arthur Conan. 1921. *The Coming of the Fairies*. New York: G.H. Doran Co.

Greaves, Hilary, and William MacAskill. 2021. "The Case for Strong Longtermism." *Global Priorities Institute Working Paper*, no. 5.

Haselager, Pim. 2024. "From Angels to Artificial Agents? AI as a Mirror for Human (Im)Perfections." *Zygon: Journal of Religion and Science*, July. https://doi.org/10.16995/zygon.11659.

Karr, Mary. 2015. *The Art of Memoir*. HarperCollins.

Kurzweil, Ray. 2024. *The Singularity Is Nearer: When We Merge with AI*. Viking.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Vallor, Shannon. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.