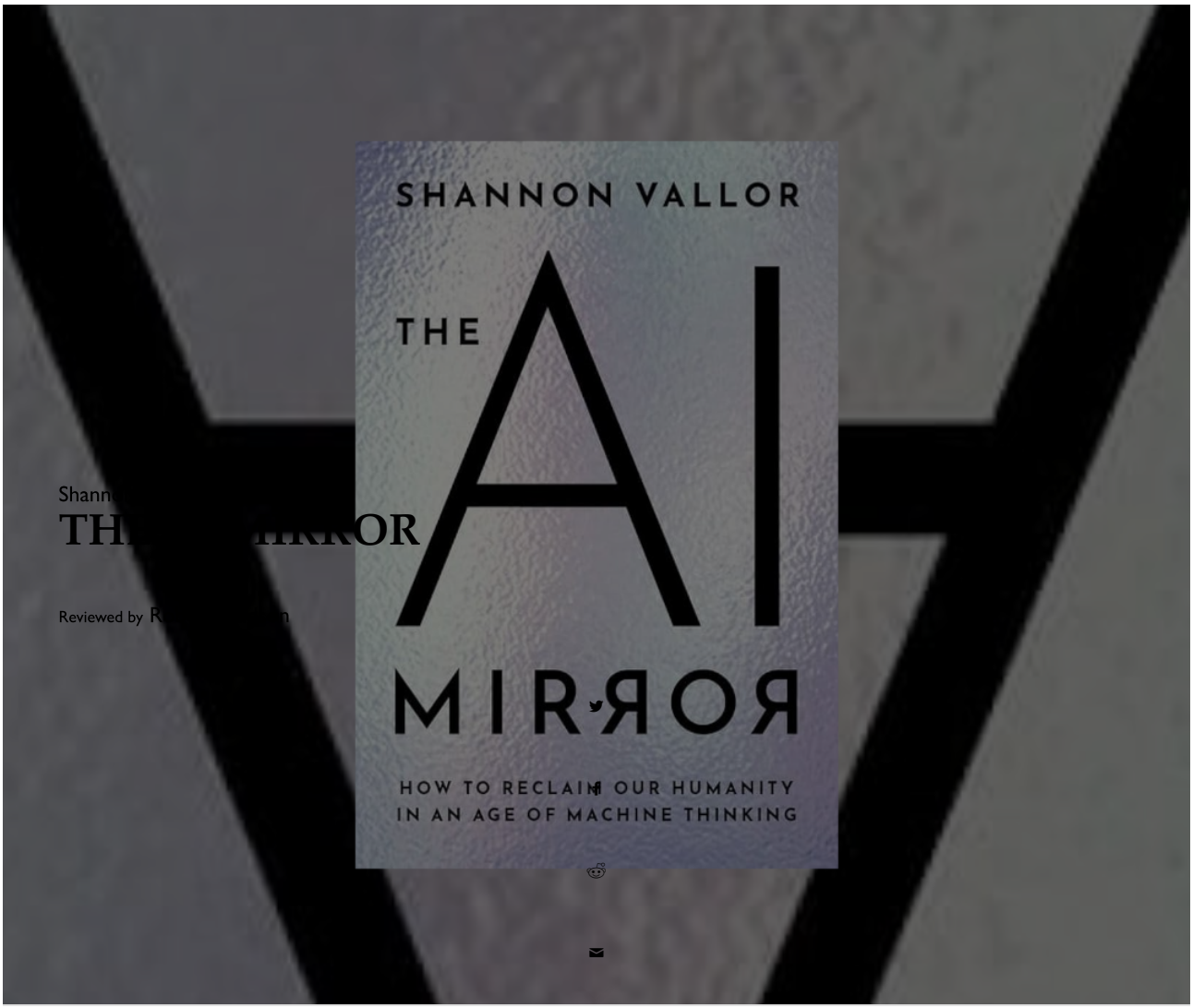

THE AI MIRROR

Shannon Vallor

Reviewed by
Robert Hudson

The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking
Shannon Vallor
Oxford: Oxford University Press, 2024, £22.99
ISBN 9780197759066

Cite as:
Hudson, R. [2025]: 'Shannon Vallor's *The AI Mirror*', *BJPS Review of Books*, **2025**, https://doi.org/10.59350/wq43t-20c50

Shannon Vallor's *The AI Mirror* is a useful introduction to the various moral, social, and political problems raised by artificial intelligence (AI). There are many things that can be said to be artificially intelligent. Vallor's focus is a machine that learns using algorithmic, statistical mathematics to 'produce novel outputs of the same general kind as their training data (such as images, sounds and sentences)' (p. 19). This sort of intelligence is analogous to the

intelligent ability of humans who, upon provision of limited input information, successfully generalize to unreported cases. For example, in learning colour language, humans need only a few cases for them to generalize successfully about colours in non-identical situations.

The artificial intelligence possessing this analogous capacity is called 'generative AI'. Generative AI models, such as OpenAI's ChatGPT, Microsoft's Copilot, Google's Gemini, and so on, are large language models (LLMs) that, upon provision of linguistic phenomena drawn from the Internet, generate novel text reflecting a broad swath of human knowledge. Vallor calls these models 'AI mirrors'. They mirror our culture and, as we are participants in this culture, they also mirror us, a phenomenon she illustrates using the motif of Narcissus from Ovid's *Metamorphoses* (pp. 4–6). Narcissus views himself reflected in a pool just as we see ourselves reflected in an informed AI model.

Vallor views the rise of LLMs with consternation. In general terms, she complains that AI lacks the experience of being human. AI doesn't feel or experience things, and doesn't understand the text it produces. How then can it generate meaningful pronouncements? Vallor also worries that AI continually makes mistakes, mistakes that often reflect a biased source of data. Crucially, for her, most of the cultural 'training data for AI models heavily overrepresent English language text and speech, as well as data from young, white, male subjects in the Northern Hemisphere, especially cheap data generated in bulk by online platforms' (pp. 48–49). These sorts of concerns fill the first two chapters of her book.

To put these concerns in context, LLMs are highly intelligent. They're programmed to reason efficiently and cogently, on the basis of an enormous data base, producing text that is often relevant and informative, something an 'intelligent' human being would do. Admittedly, LLMs are not perfect: they continuously need new information, new data, to stay relevant and informative, which they succeed in doing by being (in Norbert Weiner's terminology) 'cybernetic'. Cybernetic computational systems (computational servo-mechanisms) accomplish a particular purpose successfully over time by continually adjusting the character of their functioning in response to input data.

In this respect, the cybernetic systems with which Vallor is concerned are similar to evolutionary, biological systems. In order to survive, a biological organism modifies its behaviour on the basis of feedback from the natural environment. In the spirit of this analogy, there is a tradition in current AI theorizing that sees AI as developing in accordance with evolutionary pressures, much as biological organisms do. An early source of this tradition is Samuel Butler's ([1863]) paper, 'Darwin among the Machines', incorporated as 'The Book of the Machines' in (Butler [1872]) and cited by Vallor (pp. 70–74). More recent expressions of this tradition are found in John Holland's ([1975]) formulation of the genetic algorithm, itself developed by John Koza ([1992]). On this view, one can create a machine that automatically learns on its own, just as a biological organism automatically learns on its own, by means of natural selection, without being surveilled. As Koza et al. ([1996], p. 3) explain, the goal is to create an automatic learner who (or that) solves 'problems without being explicitly programmed, [...] without being told exactly how to do it'—here paraphrasing the words of Arthur Samuel ([1959]), the original developer of a self-learning, computational machine that could play checkers. Given the precedent set by self-learning, cybernetic biological organisms that undergo successful modifications by means of natural selection pressure without intentional design, the prospect of non-natural, generative AI systems, such as LLMs, becomes credible.

Whether constructed on the model of natural systems or not, Vallor correctly notes that predictive, generative AI is typically opaque to our understanding (p. 106). An 'opaque' learning mechanism is one where we are unable to predictively formulate, or reconstruct in a logical way, how an automatic learner solves problems. In this connection, Weiner ([1961], p. x) distinguishes between 'white' boxes and 'black' boxes, where the former are 'bodies of known structure' that statistically represent input data and the latter represent 'an as yet unanalyzed nonlinear system'. Any system is ultimately based on some black box, and in predictively capturing the results of a system we construct 'a multiple white box which [...] will automatically form itself into an operational equivalent of

[a] black box' ([1961], p. xi). This is, for Weiner, something we want if our goal is to control and manage the environment, and despite the inherent and unresolved opaqueness of the system.

For Vallor, this opaqueness is a serious obstacle for the use of generative AI. Using terminology borrowed from Wilfred Sellars, one cannot negotiate with AI in the 'space of reasons' (p. 106) because of this opaqueness. We cannot discuss with AI the nature of its programming, for AI doesn't have a conception of its own program that it can discuss with us, given its basis in a black box. Because of this lack, the result on Vallor's view is a disrespectful situation for human beings who value their ability to reason with others. For example, she cites the fictional AI 'Multivac' created by Issac Asimov whose predictive expertise renders democratic decision-making moot (p. 112). Similarly, if we relegate decision making to AI, there is an inevitable lapse of political accountability since 'opaque AI decision systems are highly attractive tools for those in power [offering] a virtually bulletproof accountability shield' (p. 119).

Whether AI is unable to occupy the space of reasons depends on our purpose in creating it. It is not inconceivable that we might create AI capable of communicating with us in the space of reasons, instead of a purely opaque and prescriptive AI. This is the case with Joseph Weizenbaum's ([1966], p. 42) pioneering ELIZA program, whose Rogerian response patterns provide interlocutors the intense phenomenon of being 'heard and understood'. What more is required for something to occupy the space of reasons than to offer the experience of being heard and understood? Originally, Sellars had required that occupying the 'logical space of reasons [involves] justifying and being able to justify what one says' (Sellars [1956]; quoted by Vallor, p. 106). This sounds distinctively like something LLMs can do with their enormous lexical databases.

In the remaining chapters of the book, Vallor's objections mostly stem from her anti-capitalist sentiments targeting a 'rising techno-theocracy' (p. 217). She scorns the leaders of the tech industry, the 'Silicon Valley billionaires' who think 'the fruitful multiplication of intelligent, benevolent machines bearing 'digital consciousnesses' might be a worthier goal for the future than sustaining a world for imperfect people' (p. 155). These 'powerful men who see themselves as a natural elite' (p. 74) endorse the 'dominant values and tendencies of the tech ecosystem—the unchecked pursuit, consolidation, and elite control of wealth and influence', values and tendencies conveniently rebranded as 'altruism' (p. 157). Even more provocatively, these leaders are 'colonizers', advocates of 'the earlier American settler myth, which saw the push to the west and colonization of Indigenous lands', and 'sought to overwrite the living values and institutions of Indigenous Americans and justify it with a religious claim of manifest destiny' (p. 218). The difference now is that 'today's AI theology seeks to overwrite our humane agency and potential with an imagined "superhuman" intelligence that renders ours into insignificance' (p. 218). Aggravating the political situation, for Vallor, is the current 'existential threat' posed by the climate crisis. AI mirrors, she claims, block 'the emergence of the new cultural visions of our relationship to technology that we need in order to survive' (p. 192).

These political aspects of AI result from how LLMs produce 'images, sounds and sentences'. In surveying the internet for information, LLMs naturally draw on old sources, containing outdated, sometimes biased opinions. AI developers know this and are challenged by the fact that good sources of information typically have copyright restrictions. Hence, contracts need to be negotiated with reputable news sources, such as OpenAI's recent content-partnership deals with '*Time* magazine, *Financial Times*, *Business Insider*-owner Axel Springer, France's *Le Monde* and Spain's PRISA Media' (as reported in *The Globe and Mail*, 20 August 2024). The task of accessing new data sources is an ongoing issue with AI. The developers of AI have no desire to be restrictive in this respect, since the quality of their product depends on the quantity of information LLMs have to work with. Vallor herself illustrates in a number of places how AI gives false or misleading results when its input is limited.

The general theme of Vallor's book is to keep a cynical eye on the promise of new AI technology. For her, developing technology for technology's sake is insufficient motivation, as there are pressing moral obligations on the agenda.

Given this, Vallor advocates 'the exercise of technomoral wisdom' (p. 171). AI models must be used for respectable ends, 'from social fairness and justice, to privacy and autonomy, to the transparency and accountability of sociotechnical systems, to democratic health and the sustainability of the planet' (p. 170). She closes the book by illustrating, using literary fiction, how AI technology can be used for worthy ends. Certainly, there is much reason to support such ventures. The obstacle here is that addressing moral issues is seldom straightforward. For example, her stark portrayal of the Silicon Valley techno-elite as villainous obscures otherwise complicated moral, social, and political issues. In general, however, her deep-felt normative concerns deserve wider consideration by scholars and the general public alike.

Robert Hudson
University of Saskatchewan
roh784@mail.usask.ca

References

Butler, S. [1863]: 'Darwin among the Machines', *The Press*, 13 June 1863, available at <ndhadeliver.natlib.govt.nz/webarchive/20210104000423/>.

Butler, S. [1872]: *Erewhon*, Mineola, NY: Dover.

Holland, J. H. [1975]: *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, Ann Arbor, MI: University of Michigan Press.

Koza, J. R. [1992]: *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, Cambridge, MA: MIT Press.

Koza, J. R., Bennett, F. H., Andre, D. and Keane, M. [1996], 'Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming', in J. S. Gero and F. Sudweeks (*eds*), *Artificial Intelligence in Design '96*, Springer: Dordrecht, pp. 151–70.

Samuel, A. [1959]: 'Some Studies in Machine Learning Using the Game of Checkers', *IBM Journal of Research and Development*, **3**, pp. 210–29.

Sellars, W. [1956]: 'Empiricism and the Philosophy of Mind', *Minnesota Studies in the Philosophy of Science*, **1**, pp. 253–329.

Weiner, N. [1961]: *Cybernetics or Control and Communication in the Animal and the Machine*, Cambridge, MA: MIT Press.

Weizenbaum. J. [1966]: 'ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine', *Communications of the ACM*, **9**, pp. 36–45.