Book Review of *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*

Wisdom Without a Mechanism: The Bootstrapping Problem in Institutional Practice

Bodhishka Bankar
Student, Dada Ramchand Bakhru Sindhu Mahavidyalaya
Kshitij Taluja
Undergraduate Researcher, Bachelor of Computer Science, Golden Gate University
Dr. Smrite Goudhaman
Global Professor of Practice, Golden Gate University

Author: Shannon Vallor
Title: *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*
Publisher: Oxford University Press
Publication year: 2024
ISBN: 978-0-19-775906-6
Pages: 272

Abstract

Shannon Vallor claims that today's leading AI systems act as mirrors, extracting patterns from records of human action and projecting them forward, and that the credible existential risk they pose is a diminishment of human moral confidence and imagination rather than machine domination. This review evaluates *The AI Mirror* from the viewpoint of organizational practice, asking what institutions that took the diagnosis seriously could act on. Its principal strength is what Vallor calls the bootstrapping problem: addressing present crises requires cultivating virtues our institutions do not currently reward, using only the virtues we already possess. Its principal unresolved issue is the gap between her structural description of AI mirrors as backward-facing and her proposal that the same systems can serve recovery and repair, a shift she illustrates but does not demonstrate. The book is significant for reframing AI governance as a question of institutional character rather than technical correction.

Keywords: artificial intelligence ethics; practical wisdom; algorithmic bias; higher education; action research

Shannon Vallor's *The AI Mirror* (2024) shows how commercial artificial intelligence primarily acts like a mirror: a system built on records of human thought and action, reflecting those patterns back with exaggerated distortions, and one we often take to be a window onto our future. One of her contributions is to shift where the credible existential risk lies. Vallor argues that the real risk is not machine domination but self-forgetting, the erosion of human moral confidence and of autofabrication, our continued effort to make ourselves anew.

This review assesses the book from the viewpoint of organizational practice and asks a single question: what could an organization actually do if it accepted Vallor's diagnosis? We do not attempt to evaluate every philosophical dimension of the book. Our concern is narrower and, we think, more consequential: the distance between a philosophical diagnosis and institutional action. From that vantage point, we argue that the book's greatest strength is also its least developed element, and that the unfinished work is an invitation rather than a flaw.

Vallor details the mirror process thoroughly. She relies on well-established examples: a widely used U.S. hospital risk assessment algorithm that allocated fewer resources to Black patients because it predicted cost rather than illness (Obermeyer et al., 2019), the recruitment tool withdrawn once it had learned to devalue candidates by gender, and the large language models termed stochastic parrots by Bender et al. (2021). She uses these examples to conclude that bias in AI systems is a people problem rather than a computational one, because such systems reflect biases already present in institutional behavior. She also provides a complementary case showing that an algorithm trained on a more diverse sample explained considerably more of the racial disparity in knee pain than standard radiologist assessments (Pierson et al., 2021). In our reading, and as an extension of Vallor's argument rather than a claim she makes herself, the same class of tool may either exacerbate an injustice through automation or document it for correction. Institutional intent is part of what separates the two outcomes, but it is not the whole of it. Training data, model design, deployment conditions, oversight arrangements and the incentives acting on the people who read the output all bear on which use a tool is put to.

The central and most innovative argument of the book is the bootstrapping problem presented in chapter six. Vallor concludes that the characteristics our institutions celebrate as excellence, such as perseverance, independent thinking and disruptive production, contributed to creating today's environmental and social crises. Meeting those crises therefore requires virtues we do not consistently display, cultivated using only the virtues we now have. Her proposed path forward is technomoral wisdom, an integration of technical and moral competence that she believes modern education has artificially separated, along with a reassessment of restraint, repair, restoration and care. For organizational leadership this is not an abstract difficulty. A governance reform is designed, approved and staffed by the same people, promotion criteria and reporting lines that produced the practice it is meant to correct, which is why AI governance so often reproduces the culture it was introduced to change.

At this juncture the tension becomes apparent. Vallor documents in detail the structural conservatism of AI mirrors, designed to identify historical patterns and extrapolate them forward, and she agrees with Birhane's (2021) description of such systems as conservative seers. Yet in her constructive chapter she suggests that those same mirrors can function as a means of healing, and she resolves the difficulty mainly through the figure of Prudence, who in Renaissance iconography holds a backward-looking mirror so that she can walk forward. The inference drawn here is that the solution is presented as an image rather than an argument. No demonstration is made at the level of design, incentive or governance of what separates the destructive use of these tools from the restorative one, and that mechanism is exactly what practitioners would need. What would count as one is reasonably clear: design principles distinguishing diagnostic from predictive deployment, procurement and incentive structures that reward the former, governance arrangements naming who reviews the output and against what standard, and documented cases in which an organization changed a policy because of what a mirror showed it. Because *The AI Mirror* is a philosophical argument rather than an empirical study, not yet demonstrated is a more accurate characterization than unsupported, and this marks the most promising area for future research.

For organizations, the book's primary practical value lies in a distinction Vallor offers. She adopts Brynjolfsson's (2022) contrast between automation and augmentation, arguing that current economic incentives favor the former. Organizations deploying AI mirrors to automate decision-making should expect, on her account, to replicate the patterns already embedded in their records. Augmentation, on this reading, changes where the system sits in the decision rather than how accurate it is. Instead of issuing a score a caseworker or recruiter is expected to accept, the tool surfaces the pattern, the exceptions and the cases it cannot account for, and a named person decides and records why. The judgment stays with someone who can be asked to justify it. Government agencies face a sharper version, since Vallor documents automated welfare fraud detection systems whose failures caused severe harm to wrongly accused families. Her contention that such tools are better aimed at corruption among the powerful than at scrutiny of the already marginalized is, in our reading, among the few directly actionable proposals in the book, and one that nonprofit and advocacy organizations are well placed to test.

The implications for action research follow immediately. Vallor's Prudence outlines the reflective cycle: examine how previous decisions actually turned out, identify the unintended effects they produced, intervene, and evaluate the result. That sequence corresponds closely to

the action research cycle of constructing the issue, planning action, taking action and evaluating it, which is repeated rather than completed (Coghlan & Brannick, 2014). Documentation is what makes the cycle usable in an institution. It would need to record what it expected before intervening, what it observed afterwards, and what it changed as a result, so that each turn is auditable and the next begins from evidence rather than recollection. An organization could operationalize this by treating its own record of past decisions as the mirror, using pattern-detection tools to surface disparities in hiring, procurement, admissions and service delivery, then asking what the pattern implies for policy rather than for prediction. Vallor develops the point through Arista's work on machine learning for Hawaiian language recovery, where success means the tool is no longer needed once the community has regained its teaching capacity. Success defined by a technology's eventual withdrawal is unusual, and worth practitioners' attention.

The argument is uncomfortable for higher education. Vallor states that hyperspecialized technical courses divorced from the humanities leave graduates responsible for consequential systems without preparation for it, and she argues that moral capacities are developed through practice, so that automating deliberation produces what she calls moral deskilling. Universities in resource-constrained or underrepresented contexts face this differently rather than uniformly, though a common pattern is the adoption of models developed elsewhere without any input into what those models encode, one manifestation of the underrepresentation in training data that Vallor documents. This maps onto a real tension in the 2030 Agenda, where the innovation and infrastructure commitments of Goal 9 can be advanced by AI deployment that simultaneously widens the inequalities addressed by Goal 10 (United Nations, 2015). A single deployment can do both at once. An institution that installs an imported diagnostic or admissions system adds measurable technological capacity, which registers against Goal 9, while the assumptions encoded in that system continue to disadvantage the same populations, which works against Goal 10. The first effect is considerably easier to report than the second. Vallor's criticism of AI for Good makes the point sharply: such initiatives typically advance the Goals while leaving untouched the structures that produced the harm being measured.

The AI Mirror will be most useful to higher education leaders, policymakers, and decision-makers in business, government and nonprofit organizations that are adopting AI systems faster than they can articulate what those systems are for. Its greatest accomplishment is to have identified the bootstrapping problem clearly enough that it can now be worked on. Read as an implementation framework for responsible AI governance, the book will disappoint, and it

makes no claim to be one. Read as a way of reframing the questions an institution asks before it adopts a system, it is genuinely useful, and that is the use we would recommend. Its limitation is that a book arguing we cannot reason our way out of an inherited pattern cannot itself supply the practice needed to break it. That task falls to institutions, and to the practitioners willing to examine their own records first.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns, 2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Brynjolfsson, E. (2022). The Turing trap: The promise and peril of human-like artificial intelligence. *Daedalus, 151*(2), 272–287. https://doi.org/10.1162/daed_a_01915

Coghlan, D., & Brannick, T. (2014). *Doing action research in your own organization* (4th ed.). SAGE Publications.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., & Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine, 27*(1), 136–140. https://doi.org/10.1038/s41591-020-01192-7

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development* (A/RES/70/1). https://sdgs.un.org/2030agenda

Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press. https://doi.org/10.1093/oso/9780197759066.001.0001