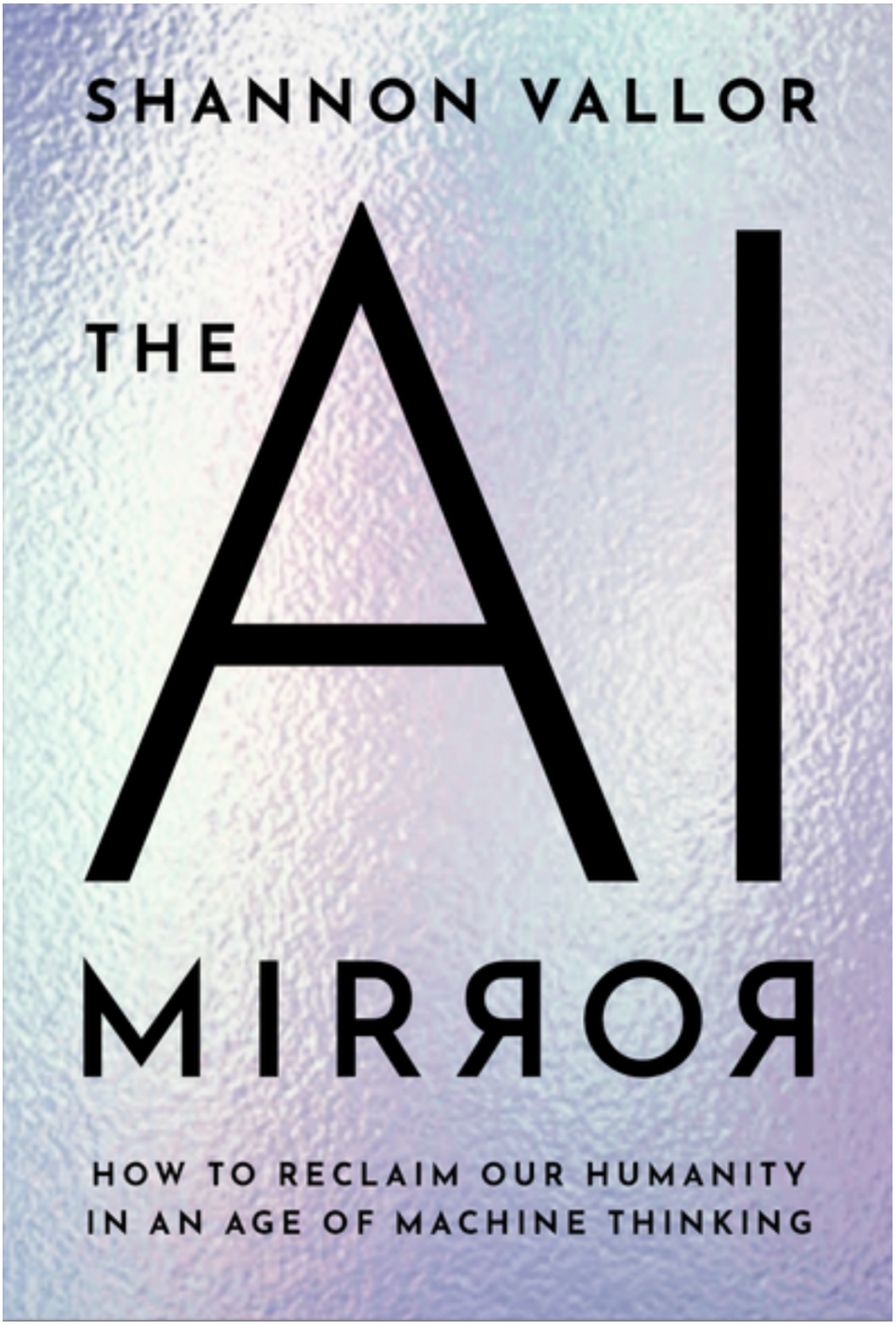
# AI as a Mirror

# AI as a Mirror

Review of S. Vallor (2024). *The AI Mirror. How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford University Press, New York, 263 p.

Dr. G.J. van Bussel



Shannon Vallor is a professor of the ethics of data and artificial intelligence (AI) at the University of Edinburgh. In this role, she explores how new technologies such as AI, robotics, and data science are reshaping human moral character, cognitive habits, and social practices. As one of today's leading philosophers of technology, she is a key figure in contemporary debates about humanity's technological future. Her 2016 book, *Technology and the Virtues*, is an attempt to rethink virtue ethics in the context of technological upheaval and breakthroughs. In it, she presents 'technomoral virtue ethics', a framework that critiques how technological breakthroughs erode traditional moral capacities, while also proposing strategies for cultivating new virtues adapted to our era of technological disruption. [i] Her latest book, *The AI Mirror* (2024), builds on this research and focuses on three interconnected themes: the moral agency of AI systems; the effects of algorithms on human self-understanding; and the conditions needed for responsible innovation. In *The AI Mirror*, Vallor provides a moral and philosophical analysis of AI and its role in shaping knowledge, values, and morality.

Nearly a year after publication, *The AI Mirror* remains relevant, its title perfectly expressing Vallor's central thesis: what we call 'AI' is not intelligence but a reflection — one that mimics human cognition while fundamentally lacking understanding. The mirror analogy proves very powerful, exposing how these systems merely process and refract historical data imbued with human conceptions (and misconceptions) including conscious or unconscious biases and misrepresentations of reality. [ii] AI serves as a mirror that exposes and intensifies societal biases, aspirations, and moral contradictions. Vallor argues that AI not only mirrors but recursively influences human cognition.

Employing the mirror metaphor, she warns that without deliberate human intervention, AI will perpetuate biases, distortions, and blind spots rather than fostering meaningful progress. These systems 'aren't designed to be accurate — they are designed to sound accurate' (p. 121.) This distinction is vital, highlighting how large language models trade in plausibility rather than truth, and how their outputs, despite their fluency, remain cognitively hollow. Vallor asserts that the self-image projected by AI is deceptive and that an uncritical fascination with AI poses significant dangers. 'It is their power to induce in us a type of self-forgetting — a selective amnesia that loosens our grip on our own human agency and clouds our self-knowledge. It is an illusion that can ensnare even the most technologically adept among us' (p. 2.)

The mirror analogy is sustained throughout the book. Each chapter opens with an epigraph on mirrors drawn from renowned writers, poets, or philosophers. The preface begins with an epigraph from E.M. Forster's 1909 short story *The Machine Stops*, which reflects on humanity's guidance by various forms of mirroring but now find themselves in an alienating relationship with machines. Vallor's work, as she describes it, is 'about our humanity and our technology. It's about how each constitutes the other, and how we have become alienated from both. It's about how to muster the courage to mend a self-inflicted wound, inflicted long ago by a philosopher's futile attempt to cleave them apart' (p. vii). Here, Vallor references Plato's exclusion of 'craftsmen', those engaged in technical arts, from political deliberation. Craftsmen were not considered to constitute a rational elite, as 'common crafts' were looked down upon (p. 190, and Chapter 6, note 12.) The preface gives a concise summary of the book's core themes, with each subsequent chapter building on these concepts.

The introduction opens with an epigraph from the rock band The Velvet Underground and centres on the question of what it means to be human. Vallor emphasizes the difference between AI and human cognition, stressing that computers lack human capacities. She draws upon the Spanish philosopher José Ortega y Gasset, who argued that the essence of humanity lies not in rationality but in our embodied existence, our 'embodied lives' (p. 12.) This becomes a cornerstone of Vallor's argument, with recurring references to Ortega y Gasset throughout the text. Vallor adopts Ortega's view that humans are perpetually engaged in autofabrication — 'the task of making ourselves' (p. 12.)

Our nature, she asserts, is future-oriented: we must 'choose to make ourselves and remake ourselves, again and again' (p. 206.) By contrast, AI is inherently retrospective. Its predictions and outputs derive entirely from historical training data, rendering it capable of reflecting only our past, never our future possibilities. 'The more we rely on them to know who we are,' Vallor warns, 'the more the fullness of our humane potential recedes from our

view' (p. 11.) In addition, she notes that the design of AI systems only encodes the values of a narrow demographic — 'a tiny subset, part of an increasingly homogenous tech monoculture' (p. 13) — yet are erroneously perceived as mirrors of universal humanity. This homogeneity aggravates the distortions Vallor attributes to the AI mirror, a reflection skewed by the limited perspectives of its creators. And, let me add, this homogeneity imposes a parochial Western technocratic worldview as universal, something Vallor doesn't mention.

In Chapter 1, Vallor offers a detailed examination of the fundamental distinctions between human beings and AI. The chapter assesses AI's potential while exploring how we might employ it in less destructive ways. Vallor introduces this chapter with an epigraph from Ovid's tale of Narcissus — the youth who, enchanted by his own reflection, becomes incapable of engaging meaningfully with life. This serves as a metaphor for contemporary humanity: we have grown infatuated with our technological creations yet struggle to reclaim agency over them. Vallor identifies this dynamic as AI's most problematic aspect: 'we surrender the task of understanding ourselves, our history, our differences, and our shared humanity to machines that merely fabricate variations on stories already told — and only by the most privileged. We're on the brink of surrendering the urgent task of engineering ourselves and our societies anew to mindless tools without hope or vision, that only predict what the historical data say we will probably do next' (p. 36.) This critique echoes her earlier warning about AI's homogenizing effects.

Sylvia Plath's 1961 poem 'The Mirror', which opens Chapter 2, articulates the function of a real mirror. This provides the basis for Vallor's defence of the mirror analogy against conceptions of AI as a form of mind, that have become increasingly prevalent in Silicon Valley instigated research. She builds on the distinctions she has recognized between humans and machines. Vallor argues that AI does not interact with its environment in the same way as minds do: 'An AI mirror is not a mind. It is a mathematical tool for extracting statistical patterns from past human-generated data and projecting these patterns forward into optimized predictions, selections, classifications, and compositions' (p. 38.) She substantiates this distinction with multiple examples, stressing that AI systems lack the qualities of mindedness. Vallor argues that AI's true nature cannot transcend unconscious biases because it lacks biological embodiment, and its outputs remain tethered to historical data. This limitation, she warns, reinforces historical biases, which in turn shape future actions, entrenching those harms more deeply. [iii]

Chapter 3, *Through the Looking Glass*, introduced by an epigraph from Lewis Carroll, builds on her examination of the human mind in Chapter 2 and examines the virtue of imagination and its precarious relationship with AI. Using the thesis of her previous book that human virtues can and should guide the development of 'wiser' technologies, [iv] Vallor contends that contemporary (science fiction) narratives of AI have trapped us in a dangerous fantasy: a 'looking glass' formed from self-referential imaginings. She argues that the seductive vision of artificial general intelligence (AGI) (the 'superhuman' view adhered to by many of Silicon Valley's elite) has distorted policy priorities, diverting attention from genuine human flourishing to the impractical and unrealistic pursuit of artificial agency in the (very far) future.

Vallor likens this to 'strong longtermism', wherein 'far future threats, even highly speculative ones like AGI, simply dwarf even the most urgent moral claims of presently living or soon-to-be-born humans' (p. 80.) Such preoccupations are misguided, she asserts, given that AI systems 'can't think their way out of a paper bag' (p. 81.) The imagined spectre of 'superhuman intelligence', Vallor warns, constitutes a threat because it is a delusion — one that recasts the mirror's limited reflections as a window to the future. These projections, however, represent only a narrow subset of possibilities, while obscuring pressing, tangible dangers in the present.

In one of her essays, Vallor contests OpenAI's characterization of AI as 'superhuman' (a term laden with implications of self-awareness and cognitive indistinguishability from humans) on the basis that AI comprises 'highly autonomous systems capable of outperforming humans in most economically valuable tasks'. I argue that this definition is not only philosophically naïve, but also ethically hazardous. Equating 'autonomous task execution' with 'superhuman' cognition commits a category error by collapsing the distinction between functional proficiency and genuine agentic capacities, such as intentionality, self-reflection, and moral reasoning. It risks perpetuating a misleading anthropomorphism, whereby optimization benchmarks are mistakenly regarded as indicators of consciousness. Therefore, Vallor's critique of AI's 'mindedness' is valid: surpassing human performance in domain-specific tasks (a feat already achieved by narrow AI) does not entail the presence of a mind. OpenAI's rhetorical shift from 'superhuman cognition' to 'economic outperformance' suggests a tacit acknowledgement that the former is impossible under current paradigms, prompting a strategic retreat to more manageable ground. [v]

Subsequent chapters build on themes introduced in earlier sections. However, they become repetitive when viewed from different perspectives using the mirror epigraphs.

In Chapter 4, Vallor builds upon her critique of AI's alleged mindedness using Wilfrid Sellars's framework of the 'logical space of reasons' (p. 106.) She argues that AI lacks the capacity for deliberative, context-sensitive reasoning that is characteristic of human cognition. The chapter's epigraph from Da Vinci illustrates this distinction, highlighting the constitutive role of reason in artistic creation, a process requiring contextual understanding and intentionality that AI cannot replicate. Vallor systematically demonstrates how this divide manifests in ethical reasoning. While human judgement employs empathy to recognize moral significance (e.g. suffering) and adapt to new situations, AI is limited by the statistical patterns in its training data. It is unable to critique its own biases or adapt its 'goals' in light of new moral insights.

While this argument successfully differentiates human reasoning from artificial processing, it risks two oversimplifications. Firstly, one might argue that even without genuine understanding, AI can support human judgement in areas such as medical diagnostics, though Vallor would probably counter that this instrumental usefulness does not constitute participation in the 'space of reasons'. Secondly, the critique focuses on the technical limitations of AI while ignoring the institutional logics (e.g. neoliberal efficiency metrics) that exploit these flaws. Crucially, neither point negates Vallor's core thesis. AI's inability to engage in reasoning creates risks when it is deployed in socially consequential domains such as criminal justice or care work, where its outputs are mistaken for legitimate judgements. Her framework could be further strengthened by engaging with techno-optimist challenges (e.g. hybrid neurosymbolic systems) without conceding ground on the ontological distinction.

Chapter 5 expands Vallor's critique by examining the ramifications of AI's incapacity to engage in what Robert Brandom terms the 'game of giving and asking for reasons', particularly where empathy is concerned. [vi] The chapter opens with Rumi's metaphor of a rust-covered mirror reflecting nothing, which serves a dual purpose: it captures both the erosion of human empathic faculties through technological mediation and the ontological void at the heart of AI systems. Vallor uses the metaphor of the 'empathy box' to analyse how AI simulates empathetic interactions (in therapeutic chatbots, care robots, and other applications) while lacking genuine understanding. This simulation, she argues, reduces empathy (a moral practice rooted in reciprocity, vulnerability, and ethical commitment) to a transactional service.

While acknowledging that such systems might offer limited utility for individuals who experience empathy deficits, Vallor issues a stringent warning against over-reliance on artificial empathy. She demonstrates how this dynamic extends capitalism's commodification of human feeling into the technological realm. Her central claim remains uncompromising: the ethical essence of empathy is diminished when reduced to algorithmic functions. It is a moral skill requiring reciprocity, vulnerability, and ethical commitment that cannot be automated.

Chapter 6, framed by McLuhan and Fiore's metaphor of 'look at the present through a rear-view mirror', [vii] traces some fundamental philosophical misconceptions about technology and the virtues back to Plato and Aristotle. This brings the reader back to the preface. The exclusion of craftsmen created a false duality between technical excellence and ethical wisdom. Vallor attempts to overcome this divide with her 'technomoral' virtue framework. Technology only makes sense when it is understood as a way of life in which attending to the needs of others is central.

In the concluding chapter, Vallor articulates her vision for ethically reoriented technology. She proposes the following:
(1) comprehensive moral education for technologists.
(2) the cultivation of 'technomoral virtues', and
(3) structural reforms in AI governance emphasizing transparency and accountability mechanisms.
While these solutions are theoretically coherent, they underestimate the material and ideological constraints of late capitalist technoculture. Although Vallor's emphasis on ethics training for engineers is laudable, it fails to adequately address the impotence of individual moral agents within corporate structures. Even engineers who are ethically conscious face institutional imperatives that privilege shareholder value over moral considerations. Furthermore, her proposed technomoral virtues encounter the 'incommensurability of traditions', the idea that

different traditions, whether philosophical, scientific, or cultural, may be so fundamentally different that they cannot be compared or evaluated using a common standard or measure. Confucian, Aristotelian, and Ubuntu ethics represent radically different conceptions of excellence. [viii] The proposed governance reforms rely on 'neoliberal responsibilization', which shifts ethical burdens onto individuals while maintaining power structures. [ix] Despite its transparency mandates, the GDPR has done little to curb Meta's surveillance economy or prevent OpenAI's exploitative data practices. Antitrust action and worker ownership models may be more effective. While Vallor's virtue-ethical approach is philosophically robust, it must confront the fact that ethics never operate in a power vacuum — a lesson that the AI industry continues to demonstrate.

The book's strengths are undeniable. Vallor's argument is philosophically nuanced, skilfully synthesizing virtue ethics, phenomenology, and critical algorithm studies into an original framework. She rigorously deploys Sellars' 'logical space of reasons' to define the fundamental limits of machine cognition. The book's structure is logically progressive, with each chapter cumulatively dismantling the illusion of AI's 'intelligence' and exposing its societal consequences.

Yet, the work's considerable merits are undermined to some extent by two structural limitations.

The first is *repetition*: core arguments about AI's incapacity for genuine reasoning or its amplification of historical biases reappear across chapters with minimal substantive advancement. While this reinforces the argument, it comes at the cost of analytical depth. Each recurrence makes the critique feel more like an obvious truth than a gradually developed idea. A more condensed text, distilling its sharpest critiques and remove redundant elaborations, could have amplified its impact.

The second limitation involves *contextual overextension*. While Vallor's engagement with classical philosophy, particularly the Aristotelian–Platonic tradition, is undoubtedly erudite, it diverts attention from the book's urgent contemporary relevance. A more disciplined focus on applied cases (such as the racial biases embedded in predictive policing algorithms or the moral ambiguities of caregiving robots) might have anchored Vallor's critique in concrete stakes, rendering it not just philosophically sound but empirically urgent.

In my opinion, while Vallor's critique of Silicon Valley's capitalist practices is substantively valid, framing it as a moral crusade risk diminish her philosophical arguments. Her characterization of technology leaders as a self-appointed 'natural elite' (p. 74) who rebrand 'the unchecked pursuit, consolidation and elite control of wealth and influence' as altruism (p. 157) is consistent with documented cases of ethical washing in Big Tech. [x] However, her provocative comparison of these figures to 'colonisers' invoking a 'settler myth' (p. 218) uses a charged historical analogy that, while rhetorically powerful, needs to be carefully qualified.

This parallel highlights how narratives of AGI marginalize alternative futures, particularly those of indigenous peoples and the Global South, in relation to humane technology. The analogy risks overextension by implying equivalency between territorial genocide and AI's cognitive dominance. Nevertheless, Vallor's argument gains traction in exposing the material consequences: the climate costs of AI infrastructure and the anti-democratic implications of 'superhuman' rhetoric reveal patterns of very real resource extraction and governance erosion. [xi] However, she does not need to overextend to realize this.

Despite these criticisms, Vallor's central thesis accomplishes what few works in AI ethics achieve: it exposes how the most profound danger lies not in AI's limitations, but in our relentless anthropomorphizing of them. While the book's occasional long-windedness may dilute certain arguments, this cannot diminish the gravity of its core warning. When we mistake AI's mirror for a mind, we risk undermining the very capacities that make autofabrication possible (moral reasoning, empathic judgement, and deliberative democracy) potentially triggering what Vallor might call 'recursive alienation'. This feedback loop sees AI's constrained outputs progressively restricting humanity's ethical imagination, a process that threatens to foreclose alternative futures before they can be conceived.

The challenge this poses is both immediate and profound. For autofabrication, humanity's ethical project of self-becoming, to retain its potential, we must resist two threats: the analytical logics inherent to AI, which flatten human complexity, and the influence of power structures that weaponize this tendency, codifying it into

instruments of domination and profit. Here, Vallor's work points toward what Bernard Stiegler called 'technics of care': practices of self-making that privilege moral contingency over algorithmic certainty, cognitive plurality over computational determinism, and creative resistance over passive acceptance. This is an urgent imperative, one that demands we reassert human agency in the face of systems that would outsource our capacity for ethical imagination itself.

Although it could have been more concise, this book comes highly recommended. If you haven't read it yet, you should. It offers a unique perspective on AI, and especially on us, as users of AI.

Only one minor point remains: Vallor references 'iRobot' on pages 71 and 76. I'm certain she isn't referring to the vacuum cleaner company but rather to *I, Robot*, Isaac Asimov's thought-provoking 1950 collection of stories defining 'The Three Laws of Robotics'. It is unusual for a philosopher of ethics to confuse 'iRobot' with *I, Robot*, particularly given Asimov's work explicitly grapples with the moral dilemmas of robotics, a core concern in AI ethics.

---

[i] S. Vallor (2016). *Technology and the virtues. A Philosophical Guide to a Future Worth Wanting*, Oxford University Press, Oxford.

[ii] Compare to: 'the archive is as it is, a construct configured, managed, and preserved according to organizational (or personal) demands and desires, with gaps as a result of appraisal and selection, and, as a consequence, presenting a social reality that is only mirroring a very simplified and distorted view of the contexts in which the records and the archive were generated.' G.J. van Bussel (2017). 'The theoretical framework for the 'Archive-As-Is'. An organization-oriented view on archives. Part II. An exploration of the 'Archive-As-Is' framework', F. Smit, A. Glaudemans, and R. Jonker (eds.), *Archives in Liquid Times*, Stichting Archiefpublicaties, 's-Gravenhage), pp. 42-71. Quotation: p. 62.

[iii] Based on: R. Benjamin (2019). *Race After Technology. Abolitionist Tools for the New Jim Code*, Polity Press, Cambridge.

[iv] Vallor (2016).

[v] 'Open AI Charter.' Online source, retrieved June 25, 2025, from: https://openai.com/charter/. See also: S. Vallor (2024). 'The danger of Superhuman AI is not what you think. The rhetoric over 'superhuman' AI implicitly erases what's most important about being human', *Noēma Magazine*, May 23. Online source, retrieved June 25, 2025, from: https://www.noemamag.com/the-danger-of-superhuman-ai-is-not-what-you-think/.

[vi] R.B. Brandom (2001). *Articulating Reasons. An Introduction to Inferentialism*, Harvard University Press, Cambridge.

[vii] M. McLuhan and Q. Fiore (1967). *The Medium is the Massage*. Allen Lane, London, pp. 74-75.

[viii] E. Oberheim, and P. Hoyningen-Huene (2025). 'The incommensurability of scientific theories', E.N. Zalta, and U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Online source, retrived June 25, 2025, from: https://plato.stanford.edu/archives/spr2025/entries/incommensurability.

[ix] J. Pyysiäinen, D. Halpin, and A. Guilfoyle (2017). 'Neoliberal governance and 'responsibilization' of agents. Reassessing the mechanisms of responsibility-shift in neoliberal discursive environments', *Distinktion. Journal of Social Theory*, Vol. 18, No. 2, pp. 215–235. Online source, retrieved June 25, 2025, from: https://doi.org/10.1080/1600910X.2017.1331858.

[x] J. Metcalf, E. Moss, dana boyd (2019). 'Owning ethics. Corporate logics, Silicon Valley, and the institutionalization of ethics', *Social Research. An International Quarterly*, Vol. 88, No. 2, pp. 449-476.

[xi] G.J. van Bussel (2024). 'AI's Power Demand', VBDS Blog, published November 25. Online source, retrieved June 30, 2025, from: https://vbds.nl/2024/12/31/ais-power-demand/ and G.J. van Bussel (2024). 'The Escalating Wateruse of AI', VBDS Blog, published December 20. Online source, retrieved June 30, 2025, from: https://vbds.nl/2024/12/31/the-escalating-water-use-of-ai/. Also: S. Zuboff (2019). *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*, Public Affairs, New York.