A Faustian fable

Is AI a Frankenstein's monster or an unintelligent parrot?

PHILIP BALL

MORAL AI
And how we get there
JANA SCHAICH BORG, WALTER SINNOTT-ARMSTRONG AND VINCENT CONITZER
304pp. Pelican. £25.

THE AI MIRROR
How to reclaim our humanity in an age of machine thinking
SHANNON VALLOR
272pp. Oxford University Press. £22.99 (US $29.99).

ROBOTS AND THE PEOPLE WHO LOVE THEM
Holding on to our humanity in an age of social robots
EVE HEROLD
256pp. St Martin's Press. £22.99 (US $27).

THE ATOMIC HUMAN
Understanding ourselves in the age of AI
NEIL D. LAWRENCE
448pp. Allen Lane. £25.

"WE ARE OURSELVES creating our own successors; we are daily adding to the beauty and delicacy of their physical organisation; we are daily giving them greater power and supplying by all sorts of ingenious contrivances that self-regulating, self-acting power which will be to them what intellect has been to the human race. In the course of ages we shall find ourselves the inferior race."

Thus Samuel Butler expressed in 1863 the popular fantasy indulged today about artificial intelligence: that we are in the process of building our future overlords and intellectual superiors. The title of his essay, "Darwin Among the Machines" (adapted in 1872 as "The Book of the Machines", a section in Butler's satire *Erewhon*), was used by the historian of technology George Dyson in 1998 for his survey of AI, in which he argued that a conscious mind will inevitably, sooner or later, emerge from our efforts to produce machine intelligence.

By some accounts that has happened already. In 2022 Blake Lemoine, an engineer at Google, claimed that his interactions with one of the company's large language models (LLMs) - an algorithm designed to produce conversational responses to human prompts and questions - constituted evidence that the machine was a conscious agent. Barely anyone in the AI industry was persuaded, and Lemoine was sacked by Google for violating confidentiality policies. But it's not hard to see why he was taken in by the eerily lifelike output generated by the algorithm. The speed with which such systems have developed in recent years has made it hard to know quite how close we are not just to the holy grail of "artificial general intelligence" (AGI) that can do any task at least as well as a human, but also to machines that are aware of their own existence.

Science and tech luminaries from Stephen Hawking to Elon Musk to Geoffrey Hinton, the AI expert and former Google researcher who helped to devise the methods behind LLMs, have been lining up to forecast our imminent eclipse by thinking machines unless we take action to prevent this. Hinton recently suggested that superintelligent machines are just twenty years or so away - and that, once they're here, "they won't need us anymore."

All this echoes the comment in 1965 by Irving John Good, an early pioneer of computing, that "the first ultraintelligent machine is the last invention that man need ever make". Alan Turing, Good's colleague as a code-cracker at Bletchley Park, echoed that view: "At some stage we should have to expect the machines to take control".

The four books under review survey the perils of AI, suggest steps to minimize them and explore what the onset of machine intelligence means for our sense of ourselves. All agree that there are real hazards, that Something Should Be Done and that we would be foolish to entrust that to the companies making this stuff. They also illustrate that the answers we find will depend on how deeply we are prepared to dig into the moral, social and philosophical challenges that AI creates.

The World AI Conference in Shanghai, 2023

While early efforts to develop AI focused largely on trying to identify and formalize the "rules of thinking", the systems we now have - used in areas ranging from scientific and medical data analysis to language translation and law - take a different approach. These algorithms are forms of "machine learning", trained to seek meaningful correlations within the data. For example, an LLM such as OpenAI's ChatGPT will recognize that, in human texts of all kinds, "happy" is much more likely to be followed by "birthday" than by "doesn't". There is no reasoning explicitly built into the system; it formulates its responses purely statistically. Such algorithms might learn to associate subtle aspects of a medical image such as an MRI scan, undetectable to the human gaze, with a specific disease.

In general the logic behind the AI's outputs - its "decisions", if you will - is invisible. It doesn't explain itself. So, when the algorithms sometimes generate obviously incorrect results, like identifying an image of a panda as a gibbon, it can be hard to know why. This opacity of the AI black box drives controversies about how much the machine really "knows". Some argue that the outputs from LLMs (which were made possible by technical advances in 2017) suggest that they must be able to represent genuine conceptual understanding in their circuitry: when they say that a house is a place people live in, they are not just constructing a statistically likely string of symbols, but in some sense know what this means. Others argue that the very construction of the task is such as to create an illusion of understanding, and that only as embodied beings moving around in a world of things, and interacting socially with others like us, can we truly understand what a house is. It is certainly strange and unsettling to see engineers so capably design systems to mimic humans by statistical means, only to then convince themselves that they are instead inventing a new kind of brain. But the field has never lacked for grandiosity.

What, then, could possibly go wrong? Let's be done with the apocalypse first. There's a good reason why this is science fiction's favourite AI trope. As with all science fiction, the narratives of Butler and now of the likes of Musk and Hinton need to be understood not as prophecies, but as mirrors to the zeitgeist. Such fantasies take root precisely for the reason that myths always do: because they express

Philip Ball is a science writer. His most recent book is How Life Works, *2024*

deep-seated anxieties. The warnings of the looming extinction of humankind by advanced AI are perhaps best viewed alongside the obsessions of the Silicon Valley tech barons with populating Mars to save our species, and with "curing death" by bogus dietary or technological means. None of this has much to do with the capabilities of existing technologies, or even the boundaries of known science; it is instead a projection of the terror of death that seems to haunt the mega-powerful.

That a superintelligent AI would necessarily be malicious is transparently a Faustian fable in the manner of *Frankenstein* - Victor too feared that his creature, if enabled to procreate, would wipe out humankind. Ah, but perhaps our extinction would merely happen by mistake, because we have failed to align the machine's objectives with our own: what the philosopher Nick Bostrom has called a perverse instantiation. In Bostrom's classic example, an all-powerful AI is instructed to maximize its output of paperclips, then proceeds to turn everything it can lay its prosthetic hands on into these objects, until the world is denuded of resources and perhaps of life.

But this scenario is carefully crafted to fulfil its rhetorical purpose. We have endowed this putative machine with superintelligence and superpowers far beyond ours, including an ability to reassemble matter atom by atom - yet somehow it remains so stupid that it doesn't realize the disassembly of the world and all of its inhabitants was not quite we had in mind. As the cognitive scientist Steven Pinker has said: "A characteristic of AI dystopias is that they project a parochial alpha-male psychology onto the concept of intelligence". None of the tech lords betrays the slightest awareness that their fear of an overbearingly powerful entity that dooms humankind through sheer indifference (or idiocy) might represent a serious self-own.

On the other hand, as the sociologist of technology Jack Stilgoe has recently pointed out, the AI apocalypse is highly convenient for tech entrepreneurs. "It is all-or-nothing, absolving innovators from having to engage with the messy inequities that are produced by their technologies", he wrote in *Science*. They can look concerned and grave as they debate the lurid, headline-attracting "existential risk" issue at AI summits while dragging their feet about the real, more mundane but nonetheless unsettling dangers that the technology already presents. "If we build regulations around a future fantasy", Stilgoe said, "we lose sight of where the real power lies and give up on the hard work of governing the technology in front of us."

> **The question is about the kind of society that decides to substitute technological bullshit for human relations**

In *The AI Mirror*, Shannon Vallor, an AI ethicist at Google in 2018-20 and now director of the Centre for Technomoral Futures at the University of Edinburgh, dismisses the AI apocalypse as a stale recapitulation of Butler. "There is no scientific ground for this fear", she tersely remarks. And although, in *Moral AI*, the sociologist Jana Schaich Borg, the ethicist Walter Sinnott-Armstrong and the computer scientist Vincent Conitzer entertain the need for "a containment strategy" for "superintelligent AI", they move on swiftly to more immediate and tangible concerns.

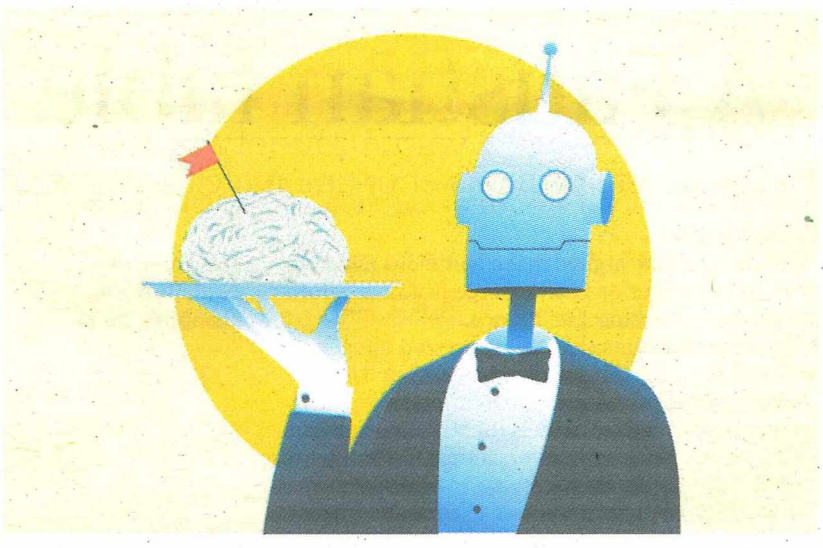
We have a tendency to trust machines as unbiased, objective and infallible, but they still get things wrong, as the Horizon Post Office scandal reminds us. Using the formidable capabilities of AI for data analysis requires skill and judgement, but off-the-shelf packages give a false sense of ease. Some fear that the scientific literature is becoming polluted with claims that cannot be replicated and are probably wrong because of misused AI: one study exposed an algorithm that confidently pronounced diagnoses from medical chest x-rays even when shown only blank regions of the images. Such errors could be a matter of life and death; mistakes made by driverless cars already have been, while in 2007 an AI-controlled anti-aircraft gun went haywire and killed nine people in South Africa.

Asking Jeeves

Robots and AI only seem human because we treat them that way

SIMONE GUBLER

ANIMALS, ROBOTS, GODS
Adventures in the moral imagination
WEBB KEANE
192pp. Allen Lane. £20.



IF ROBOTS, chatbots and cyborgs seem to have hidden purposes, "it is because their design prompts the users' deeply rooted intuitions about other beings. Like carved deities endowed with eyes, it can be hard not to conclude they have depths." So writes Webb Keane, an anthropologist undertaking "adventures in the moral imagination" in *Animals, Robots, Gods*. In this book he weaves his own reflections into a tapestry of stories drawn from anthropological fieldwork. These stories illustrate a variety of human responses to apparent ethical limit cases, cases involving the hinterlands between life and death, human and animal, and human and machine.

Although his reflections often contain contentious claims (that moral justification is context-sensitive, that contemporary moral philosophers fail to grapple adequately with moral diversity for sociological reasons), Keane largely avoids arguing for them - an arch tendency liable to frustrate some readers. Perhaps, however, this is a deliberate stylistic choice: a choice to show rather than tell. After all, much of the book's charm comes from its quicksilver turns, deft transitions and momentary flashes of family resemblance amid a rich array of ways of life. Among Keane's subjects are a person who dates a chatbot, another who confronts cancer as karmic punishment and a village that must struggle with the hungry ghosts of those who died without progeny. Taken together these cases serve to illustrate that ethical practice is highly contextual and that people in varied environments will engage in creative forms of negotiation at the putative limits of ethical concern. While the norms adopted vary according to context, their authors share some common features. One is a creative facility with ethics in a world that repeatedly pushes us to renegotiate the scope of ethical concern. Another is a drive to posit divinity. It is in the boundary lands of ethics, Keane suggests, that we often encounter (or create) our gods.

This latter tendency becomes important in the final and most substantial part of the book, where the author addresses our budding relationship to artificial intelligence. Many people see large language models (LLMs) as promising candidates for "artificial general intelligence" - as being able to think like us in key respects (something evidenced by capacities for conversation, experiential learning and performing open-ended tasks). Against this tendency, Keane is sympathetic to the idea that they are "stochastic parrots": although able to generate apparently meaningful language, LLMs fail to understand its meaning. They fail to *think* as we do.

The author moves to reassure those who think we are encountering novel subjects of ethical concern, and those worried that AI will ultimately claim ascendancy over humanity. He counsels against falling into a familiar pattern of "moral panic". Although LLMs look new, we need to remember that human beings have been negotiating comparable ethical boundaries for quite some time. Keane also reminds us that, when confronted by challenges at the margins of ethical life, we tend to project our subjectivity onto the targets of concern, then sacralize it. Robots and AI "only seem human because we actively participate in treating them this way. As they become powerful, they can start to seem superhuman ... But their divinity too is due to human collaboration ... the ways we interact with near-humans and superhumans are not new". It is we who ultimately determine how AI devices operate and what they mean. Now, if true, this would be *comforting. But alas, it might well be untrue.*

First, it's worth noting that it is controversial among programmers and theorists of AI to insist that LLMs are, or will remain, stochastic parrots. Second, while the sorts of questions we ask to determine the ethical status of AI are familiar (is it conscious, does it feel as we do, what rational capacities does it possess, is it capable of exercising moral agency?), the subject of these questions is novel - we have never encountered LLMs before. And AI is not only novel, but also increasingly ubiquitous and powerful. Many reasonable people working in tech have concluded that it poses a profound and imminent threat to all human ways of life.

In discussing our tendency to project godlike properties onto machines, Keane examines the decision to name an early web-search service Ask Jeeves. He describes the literary Jeeves (P. G. Wodehouse's famous valet) as a character who seems "to have no ego of his own", who "cannot be insulted", as someone who "exists only to perfectly anticipate and fulfil the needs of his upper-class employer". Are the makers of programs such as Ask Jeeves designing human-seeming things to be subject to relations of domination? Probably. But it's also interesting to note that the characteristics Keane invokes aren't the characteristics of the literary Jeeves at all - as anyone who recalls the matter of the white mess jacket will immediately appreciate.

> "I fear that you inadvertently left Cannes in the possession of a coat belonging to some other gentleman, sir."
> I switched on the steely a bit more.
> "No, Jeeves," I said, in a level tone, "the object under advisement is mine. I bought it out there."
> "You wore it, sir?"
> "Every night."
> "But surely you are not proposing to wear it in England, sir?"

But Jeeves's amiable employer, Bertie Wooster, does plan to wear the jacket in England, and so the march to its inevitable destruction begins with all the certainty of a Greek tragedy. It's instructive that Webb Keane so sorely underestimates Jeeves, because a persistent worry about AI, and one to which he pays too little credence, is that we will underestimate it. At some point, where we will see an ego-less and willing servant, or another idol with carved eyes, there may in fact dwell a superior intellect, outfoxing us as Jeeves does Wooster, denying us the little pleasures that make life worthwhile, whether they be jackets de rigueur or free agency. ■

Simone Gubler is Assistant Professor of Philosophy at Brown University

AI will perpetuate and even exacerbate unacknowledged biases in data sets. Borg and her colleagues cite an algorithm used to predict the need for high-risk care in patients, which prioritized white over black patients with identical levels of illness because of undiagnosed biases in the data used to train it. Sometimes those biases are subtle and hard to spot. Other times, not so much: Microsoft's Tay chatbot, released on Twitter in March 2016, was transformed, in Eve Herold's words, into "a Hitler-loving, foul-mouthed sociopath" by the human input that it "learnt" from. While there are grounds for thinking that AI can in some cases help to reduce human bias, perhaps it's sometimes better for those biases be out in the open, rather than trusting the machines to transcend them.

AI creates new privacy challenges. Borg et al reveal that our personal data is being harvested, often without any meaningful consent, ever more relentlessly on the premiss that AI will enable it to be monetized, while alleged privacy safeguards such as anonymization can often be undermined by AI itself. The cautionary precedent here is not Butler's, but E. M. Forster's short story "The Machine Stops" (1909), which portrays a future where our wellbeing has become totally dependent on a great Machine that is awarded godlike status. The cybercrime attacks on the British Library and on UK hospitals have reminded us of those vulnerabilities. Borg et al explain that solutions to such problems are more than a matter of tweaking the algorithms. They explore the institutional and organizational barriers that can hinder even with the best of intentions – moral questions, for example, tend to become another department's responsibility.

The ethical issues are complicated further when AI is housed in robots, as Herold explains in *Robots and the People Who Love Them*. A science communicator and director of policy research and education for the NGO the Healthspan Action Coalition, she considers the prospects for robots in scenarios where the human-machine interface is paramount: in education and childcare, care of the elderly, sex, therapy and companionship, and the military. In all cases there are arguments for potential benefits, but also big risks, many of which stem from the ease with which we can be induced to attribute empathy, personhood and sentience where it does not reside. Give a robot even rudimentary facial expressions and we start instinctively to mirror them. This places users in danger of emotional exploitation, Herold says, by "a machine that promises connection but can never truly deliver". On current evidence, however, social robots are likely to remain a niche market compared to the pervasive influence of AI more generally; military uses are perhaps a chilling exception.

There are no quick fixes. When Borg et al rightly say that the emotional intelligence central to the development of moral AI has not been within the conventional purview of "engineers, data scientists, and AI specialists", the answer is not just to give those groups more ethical training. They are not representative of society as a whole – not in terms of sex, race, socioeconomic status or psychological profile. As a result, Vallor says in *The AI Mirror*, "the values embedded in today's AI tools are very poor mirrors of what humans as a whole want or care about". And while STEM training becomes ever more divorced from the liberal arts and humanities, "AI researchers and developers are set up for failure".

Vallor asserts that this technology "won't enable a sustainable future without reformed political institutions and economic incentives". Worse, we could end up deferring deep structural and social change in the naive hope that AI itself will create technofixes, like all the AI-identified green materials that will allegedly solve the climate crisis.

The problem is not just that AI is limited in what it can do, but that it often offers the wrong kind of solution. Herold enumerates the potential savings in childcare costs from robot nannies, but also quotes the sociologist Sherry Turkle: "What are we saying to children about their importance to us when we're willing to outsource their care to a robot?" AI can find uses in education, but not as a substitute for human interaction. Here and elsewhere, says Vallor, we will all be the losers from "the systematic replacement of reflective discernment with mindless prediction".

At root the problem is that, by design, AI presents a veneer of omniscience, authority, objectivity and even wisdom, while that same design in fact precludes any such qualities. When AI "gets it wrong" – when it supplies an unconvincing, off-topic response or identifies a gibbon as a panda – we should remember that the algorithm hasn't actually malfunctioned. As far as human interactions are concerned, all its responses are, to use the technical term, bullshit. (The word has a sizeable entry in Vallor's index.) That doesn't make it inevitably devoid of value – as Herold shows, in some situations a robot carer or teacher might be better than none – but in the end the question is about the kind of society that decides to substitute technological bullshit for human relations.

Where does that leave us? This is the question Neil Lawrence, a professor of machine learning at the University of Cambridge, takes on in *The Atomic Human*. He should be well placed to do so, as a former director of machine learning at Amazon who has also consulted for Facebook AI. (His Cambridge position is supported by the AI company DeepMind.) On the other hand, one might wonder if this means he has drunk the Kool-Aid. "Like Cicero's notion of art, music and literature cultivating our minds, Amazon cultivates its employees through training them in the company's priorities", he writes, like Jeff Bezos channelling Alan Partridge. One also has to wonder how far to trust someone who realized, by his own admission, that Dave Eggers's novel *The Circle* (2013) was a dystopian parody, and not a straightforward fictionalization of a social media company, only after he finished the book and read the cover notes.

The Atomic Human makes some sound points, but you might need help from AI to extract them. We get long digressions about Erwin Rommel's North African campaign, the Enigma machine, radar (Lawrence is very keen on war stories), Lewis Carroll, hydrodynamics, the trial of Socrates, etymology and weather forecasting, sometimes leaping from one to the other in successive paragraphs, and none trusting the reader to have even the most basic knowledge of any of them – Lawrence can't use terms such as "gaslighting" or "Orwellian" without giving us plot summaries of their sources. These stories are presumably analogies, but if an analogy requires a detailed historical retelling, then it is not doing its job.

It's a shame, because he understands the problem: in a nutshell, "we can't leave it to big tech if we want AI to be deployed for the wider benefit of society ... It is only by bringing different voices and perspectives together that we can provide good steerage". And he too dispatches notions of super-intelligence and "technological singularities" as "hooey", saying that intelligence is not simply some sauce that entities have more or less of.

His "atomic human" is the indivisible "kernel of humanity" that remains when we have outsourced all other skills to the machine. His suggestions for what it contains – factors such as empathy, intuition, culture – are reasonable enough, but this seems a strange way to frame the issue. AI is fundamentally distinct from the human mind in the way it works. Systems such as LLMs are ever more convincing mimics of human behaviour, engineered to hoodwink so effectively that even experts like Hinton are misled by it.

"Artificial general intelligence" now serves the purpose that AI itself once did: to provide a goal that can be said to be just around the corner precisely because we bothered neither to define it nor to consider carefully enough what it entails. Lawrence quotes Frank Rosenblatt, a pioneer of the early science behind deep learning, as saying in 1958 that we'd soon have artificial brains "that could reproduce themselves on an assembly line and which would be conscious of their existence", which indicates that he had little idea what he was talking about.

Just as the pinnacle of intelligence for Rosenblatt's AI peers was an ability to play chess, so today AGI seems to amount to little more than adding human-parroting conversation and pastiche music composition to the repertoire. It's hardly a general intelligence worth wanting. The real point is to ask not how we differ from AI, but why the differences matter. Thus, "human-level AGI" is not so much a technical concept as an invitation to reinvent ourselves in the machine image. That idea pervades the AI field, and always has done. Hinton has suggested that training an AI with human feedback is like "parenting for a supernaturally precocious child" – which, as Vallor says, presents a grim view of "parenting as value alignment".

In such ways, she says, AI "threatens us from within our humanity". The mechanistic and computational narratives that have prevailed in biology and neuroscience, making us soft, squishy information processors, demand that we accept a crude vision of ourselves in preparation to hand over our power to machines. "What do you do when you are a faulty machine?", Shannon Vallor asks. "Look for a better one." ■

> 'There was a vague anxiety in the air, born of a new understanding. They walked softly and spoke in whispers. The waiters served strawberries and cream.'