# 4

# Cultural Foundations for Conserving Human Capacities in an Era of Generative Artificial Intelligence

## *Toward a Philosophico-Literary Critique of Simulation*

*Frank Pasquale*

Within a few years, machine-written language may become "the norm and human-written prose the exception" (Kirschenbaum 2023).[1] Generative Artificial Intelligence is now poised to create profiles on social media sites and post far more than any human can – perhaps by orders of magnitude.[2] Unscrupulous academics and public relations firms may use article-generating and -submitting artificial intelligence (AI) to spam journals and journalists. The science fiction magazine *Clarkesworld* closed down its open submission window in 2023 because of a deluge of content likely created by generative AI. There is already evidence of the weaponization of social media, and AI promises to supercharge it (Jankowicz 2020; Singer 2018).

AI is also poised to play a dramatically more intimate and important role in parasocial and social relationships, displacing human influencers, entertainers, friends, and partners. Not only is technology becoming more capable of simulating human thought, will, and emotional response, but it is doing so at an inhuman pace. A mere human manipulator can only learn from a limited number of encounters and resources; algorithms can develop methods of manipulation at scale,

[1] "Last June, a tweaked version of GPT-J, an open-source model, was patched into the anonymous message board 4chan and posted 15,000 largely toxic messages in 24 hours. . . . What if . . . millions or billions of such posts every single day [began] flooding the open internet, commingling with search results, spreading across social-media platforms, infiltrating Wikipedia entries, and, above all, providing fodder to be mined for future generations of machine-learning systems? . . . We may quickly find ourselves facing a textpocalypse, where machine-written language becomes the norm and human-written prose the exception" (Kirschenbaum 2023).

[2] LLMs coupled with machine vision programs to evade CAPTCHAs (Completely Automated Public Turing tests to tell Computers and Humans Apart) are the specific disinformation, misinformation, and demoralization threat anticipated here. While disinformation and misinformation involve the weaponization of false information, demoralization of a group or polity can arise when its members or citizens are bombarded by one-sided narratives (of any level of truth) designed to instill shame and doubt about the group or polity, particularly when such narratives are sponsored by authoritarian regimes or extremists who severely limit or eliminate the exposure of their own subjects or followers to similarly demoralizing narratives. This theory of demoralization builds on the account of asymmetrical openness to argument which I developed in earlier work (Pasquale 2018).

based on the data of millions. This again affords computation, and those in control of its most advanced methods and widespread deployments, an outsized role in shaping future events, preferences, and values.

Despite such clear and present dangers, many fiction and non-fiction works gloss over the problem of artificial intelligence overpowering natural thought, feeling, and insight. They instead present robots (and even operating systems and large language models) as sympathetic and vulnerable, deserving rights and respect now accorded to humans.[3] Questioning such media representations of AI is a first step toward achieving the cultural commitments and sensibilities that will be necessary to conserve human capacities amidst the growing influence of what Lyotard (1992) deemed "the inhuman": systems that presume and promote the separability of the body from memory, will, and emotion. What must be avoided is a drift toward an evolutionary environment where individual decisions to overvalue, over-empower, and overuse AI advance machinic and algorithmic modes of thought to the point that distinctively human and non-algorithmic values are marginalized. Literature and film can help us avoid this drift by structuring imaginative experiences which vividly crystallize and arrestingly illuminate the natural tendencies of individual decisions.[4]

I begin the argument in Section 4.1 by articulating how Rachel Cusk's (2017) novel *Transit* and Maria Schrader's film *I'm Your Man* suggest a range of ways to regard emerging AIs which simulate human expression. Each sympathetically describe a man and woman (respectively) comforted and intrigued by AI communications. Yet each work leaves no doubt that the AI and robotics it treats have done much to create the conditions of alienation and loneliness they promise to cure. Section 4.2 examines the long-term implications of such alienation, exploring works that attempt to function as a "self-preventing prophecy": Hari Kunzru's (2020) *Red Pill* and Lisa Joy and Jonathan Nolan's *Westworld*. Section 4.3 concludes with reflections on the politico-economic context of professed emotional attachments to AI and robotics.

Before diving into the argument, one prefatory note is in order. The sections that follow touch upon a wide range of cultural artefacts. There are spoilers, so if you intend to read, view, or listen to one of the works discussed, without being forewarned of some critical plot twist or character development, it may be wise to stop reading when it is mentioned. Unlike computers, we cannot simply delete

[3] For a review article compiling many non-fiction works that either reflect or document such sentiments, see Jamie Harris and Jacy Reese Anthis (2021). Novelists may also seek to cultivate such sentiments; see, e.g., Kazuo Ishiguro's (2021) *Klara and the Sun*.

[4] As James Boyd White (1989, 2016) has argued, "A literary text is not a string of propositions, but a structured experience of the imagination, and it should be talked about in a way that reflects its character." A "structured experience of the imagination" does not offer us propositional truths about the world. However, it gives us a sense of what it means to "live forwards" (in Kierkegaard's formulation) even as we understand backwards.

the spoiler from memory, and natural processes of human forgetting are notoriously unpredictable.

## 4.1 CURING OR CAPITALIZING UPON ALIENATION?

At the beginning of Rachel Cusk's (2017) novel, *Transit*, the narrator opens a scam email from an astrologer, or from an algorithm imitating one. The narrator describes a richly detailed, importuning missive, full of simulated sentiment. "She could sense ... that I had lost my way in life, that I sometimes struggled to find meaning in my present circumstances and to feel hope for what was to come; she felt a strong personal connection between us," (2) the narrator relates. "What the planets offer, she said, is nothing less than the chance to regain faith in the grandeur of the human: how much more dignity and honor, how much kindness and responsibility and respect, would we bring to our dealings with one another if we believed that each and every one of us had a cosmic importance?" (2).

It's a humane sentiment, both humbling and empowering, like much else in the email. Cusk's narrator deftly summarizes the email, rather than quoting it, giving an initial impression of the narrator's identification with its message and author. So how did Cusk's narrator divine the scam? After relating its contents, the narrator states that "It seemed possible that the same computer algorithms that had generated this email had also generated the astrologer herself: her phrases were too characterful, and the note of character was repeated too often; she was too obviously based on a human type to be, herself, human" (3).[5] The astrologer-algorithm's obvious failure is an indirect acknowledgement of the author's anxieties: what if her own fictions turn out to be too characterful? Carefully avoiding that, and many other vices, Cusk, in *Transit* (and the two other novels in her *Outline* trilogy), presents characters who are strange or unpredictable enough to surprise or enlighten us, to respond to tense scenarios with weakness or strength and to look back on themselves with defensiveness, insight, and all manner of other fusions of cognition and affect, judgement, and feeling.

One facet of Cusk's genius is to invite readers to contemplate the oft-thin line between compassion and deception, comfort and folly. The narrator finds the algorithmic astrologer impersonator hackish but, almost as if to check herself, immediately relates the views of a friend who found solace in mechanical expressions of concern:

> A friend of mine, depressed in the wake of his divorce, had recently admitted that he often felt moved to tears by the concern for his health and well-being expressed in the phraseology of adverts and food packaging, and by the automated voices on trains and buses, apparently anxious that he might miss his stop; he actually felt

[5] Note that computer science researchers are now seeking to detect LLM-generated text via characteristics like "burstiness" (Rogers 2022; Tian 2023).

> something akin to love, he said, for the female voice that guided him while he was driving his car, so much more devotedly than his wife ever had. There has been a great harvest, he said, of language and information from life, and it may have become the case that the faux-human was growing more substantial and more relational than the original, that there was more tenderness to be had from a machine than from one's fellow man. (3)

Cusk's invocation of an "oceanic" chorus calls to mind Freud's discussion of the "oceanic feeling" in *Civilization and Its Discontents* – or, more precisely, his naturalization of Romain Rolland's metaphysical characterization of a yearned-for "oceanic feeling" of bondedness and unity with all humanity. For Freud, such a feeling is an outgrowth of infantile narcissism, an enduring desire for the boundless protection of the good parent.[6]

Marking the importance of this oceanic metaphor in both style as well as substance, Cusk's story of the astrologer's letter has a tidal structure. Like an uplifting wave, the letter sweeps us up into reflections on fate and belief. And, like any wave, it eventually crashes down to earth, suddenly undercut by the revelation that insights once appraised as mystical or compassionate are mere fabrications of a bot. Then another rising wave of sentiment appears, wiser and more distant, calling on readers to reflect on whether they have discounted the value of bot language too quickly. The speaker is vulnerable and thoughtful: someone "depressed in the wake of his divorce," who acknowledges that the very idea of a diffuse "oceanic chorus" of algorithmically arranged concern is "maddening" (3).

Rather than crashing, this subtler, second plea for the value of the algorithmic recedes. Cusk does not leave us rolling our eyes at this junk email. She welcomes a voice in the novel that, in a sincere if misguided way, submits to an algorithmic flow of communication, embracing corporate communication strategy as concern. Cusk refuses to dismiss the idea, or to bluntly depict it as a symptom of some pathological misapprehension of the world. Her patience is reminiscent of Sarah Manguso's (2018) apothegm: "Instead of pathologizing every human quirk, we should say: By the grace of this behaviour, this individual has found it possible to continue" (44). Weighed down by depression, savaged by loneliness, a person may well seek scraps of solace wherever they appear. There are even now persons who profess to love robots (Danaher and Macarthur 2017; Levy 2008) or treat them with the respect due to a human. Indeed, a one-time Google engineer recently expressed his belief that a large language model offered such eerily human responses to queries that it might be sentient (Christian 2022; Tangermann 2022).

And yet there is a clue in the novel of how a Freudian hermeneutic of suspicion may be far more appropriate than a Rollandian hermeneutic of charity when interpreting whatever oceanic feeling may be afforded by bot language. Cusk includes a self-incriminating note in the divorcee's earnest endorsement of the

[6] For an account critically engaging with this diagnosis, see William B. Parsons (1998).

"oceanic chorus" of machines: the casual contrast, and implicit demand, in the phrase "he actually felt something akin to love, he said, for the female voice that guided him while he was driving his car, so much more devotedly than his wife ever had" (Cusk 2017, 3). A robotic voice can always sound kind, patient, devoted, or servile – whatever its controller wants from it. As the film *Megan* depicts, affective computing embedded in robotics will have a remarkable capacity for rapidly pivoting and refining its emotional appeals. It is not realistic to expect such relentless, data-informed support from a person, even a parent, let alone a life partner. Yet the more robotic and AI "affirmations" are taken to be sincere and meaningful, the more human deviation from such scripts will seem suspect. Like the Uber driver constantly graded against the Platonic ideal of a perfect 5-star trip, persons will be expected to mimic the machines' perpetual affability, availability, and affirmation, whatever their actual emotional states and situational judgements.

For a behaviourist, this is no problem: what is the difference between the outward signs of kindness and patience and such virtues themselves? This is perhaps one reason why John Danaher (2020, 2023) has proposed "ethical behaviourism" as a mode of "welcoming robots into the moral circle". In this framework, there is little difference between the given and the made, the simulated and the authentic. Danaher proposes that:

1. If a robot is roughly performatively equivalent to another entity whom, it is widely agreed, has significant moral status, then it is right and proper to afford the robot that same status.
2. Robots can be roughly performatively equivalent to other entities whom, it is widely agreed, have significant moral status.
3. Therefore, it can be right and proper to afford robots significant moral status (Danaher 2020, 2026).

The qualifier "can" in the last line may be doing a lot of work here, denoting ample moral space to reject robots' moral status. And yet it still seems wise to resist any attempts to blur the boundary between persons and things. The value of so much of what persons do is inextricably intertwined with their free choice to do it. Robots and AI are, by contrast, programmed. The idea of a programmed friend is as oxymoronic as that of a paid friend. Perhaps some forms of coded randomization could simulate free choice via AI. But they must be strictly limited. If robots were to truly possess something like the deep free will that is a prerogative of humans – the ability to question and reconfigure any optimization function they were originally programmed with – they would be far too dangerous to permit. They would pose all the threats now presented by malevolent humans but would not be subject to the types of deterrence honed in centuries of criminal law based on human behaviour (and even now very poorly adapted to corporations).

Unconvincing in their efforts to characterize robots as moral agents, behaviourists might then try to characterize robots and AI as moral patients, like a baby or

harmless animal which deserves our regard and support. Nevertheless, the programming problem still holds: a robotic doll that cries to, say, demand a battery recharge, could be programmed not to do so; indeed, it could just as plausibly convey anticipated pleasure at the "rest" afforded by time spent switched off. For such entities, emotion and communication have in *stricto sensu* no meaning whatsoever. Their "expression" is operational, functional, or, in Dan Burk's (2025) apt characterization, "asemic" (189).

To be sure, humans are all to some extent "programmed" by their families, culture, workplaces, and other institutions. Free will is never absolute. But a critical part of human autonomy consists in the ability to reflect upon and revise such values, commitments, and habits, based on the sensations, thoughts, and texts that are respectively felt, developed, and interpreted through life. The ethical behaviourist may, in turn, point out that a robot equipped with a connection to ChatGPT's servers may be able to "process" millions more texts than a human could read in several lifetimes, and say or write texts that we would frequently accept as evidence of thought in humans. Nevertheless, the lack of sensation motivating both perception and affect remains, and it is hard to imagine a transducer capable of overcoming it (Pasquale 2002). More importantly, robot "thoughts" produced via current generative AI are far from human ones, as they are mere next-word or next-pixel predictions.

Consider also the untoward implications of ethical behaviourism if persons and polities try to back their professed moral regard for robots and AIs with concrete ethical decisions and commitments of resources. If a driver must choose between running over a robot and a child, should they really worry about choosing the former? (Birhane et al. 2024). If behaviour, including speech, is all that matters, are humans under some moral obligation to promote "self-reports" or other evidence of well-being by AI and robots? In some accelerationist and transhumanist circles, the ultimate purpose and destiny of humans is to "populate" galaxies with as many "happy" simulations or emulations of human minds as possible.[7] On this utilitarian framework, what matters is happiness, as verified behaviouristically: if a machine "says" it is happy, we are to take it at its word. But such a teleology is widely recognized as absurd, especially given the pressing problems now confronting so many persons on earth.

While often portrayed as a cosmopolitan openness to the value of computers and AI, the embrace of robots as deserving of moral regard is more accurately styled as

[7] See Emile Torres (2023) describing and critiquing long-termists' projection that "humanity can theoretically exist on Earth for another 1 billion years, and if we spread into space, we could persist for at least 10^40 years (that's a 1 followed by 40 zeros). More mind-blowing was the possibility of these future people living in vast computer simulations running on planet-sized computers spread throughout the accessible cosmos, an idea that [philosopher Nick] Bostrom developed in 2003. The more people who exist in this 'Matrix'-like future, the more happiness there could be; and the more happiness, the better the universe will become." See also Jonathan Taplin (2023).

part of a suite of ideologies legitimating radical and controversial societal reordering. As Timnit Gebru and Emile Torres (2024) have explained, there is a close connection between Silicon Valley's accelerationist visions and a bundle of ideologies (Transhumanism, Extropianism, Singularitarianism, Cosmism, Rationalism, Effective Altruism, and Longtermism) which they abbreviate as TESCREAL. Once ideologies like transhumanism and singularitarianism have breached the boundary between persons' and computers' well-being (again assuming that the idea of computer well-being makes any more sense than, say, toaster well-being), long-term policy may well include and prioritize the development of particularly powerful and prevalent computation (such as "artificial general intelligence" or "superintelligence") over human well-being, just as some humans are inevitably helped more than others by any given policy. An abstract utilitarian meta-ethical stance, already far more open to wildly variant futures than more grounded virtue-oriented, natural law, and deontological approaches, becomes completely open-ended once the welfare of humans fails to be the fixed point of its individualistic, maximizing, consequentialism.

Ethical behaviourism also reflects a rather naïve political economy of AI and robotics.

A GPS system's simulation of kindness is far less a mechanization of compassion (if such a conversion of human emotion into mechanical action can even be imagined), than a corporate calculation to instil brand loyalty. Perhaps humans can learn something from emotion AI designed to soothe, support, and entertain.[8] But the more such emotional states or manners are faked or forced, the more they become an operational mode of navigating the world, rather than an expression of one's own feelings. Skill degradation is one predictable consequence of many forms of automation; pilots, for example, may forget how to fly a plane manually if they over rely on autopilot. Skill degradation in the realm of feeling, or articulating one's feelings, is a troubling fate, foreshadowing a mechanization of selfhood, outsourced to the algorithms that tell a person what or how to feel (Pasquale 2015). Allison Pugh expertly anticipates the danger of efforts to automate both emotional and connective labour, given the sense of meaning and dignity that such work confers on both givers and receivers of care and concern (Pugh 2024).

The entertaining and intellectually stimulating German film *I'm Your Man* (2021), directed by Maria Schrader, explores themes of authentic and programmed feeling as its protagonist (an archaeologist named Emma) questions the blandishments of the handsome robotic companion (Tom) whom she agrees to "test out" for a firm. Tom can "converse" with her about her work, anticipate her needs and

[8] Critical data for today's affective computing arose in part from efforts to classify human emotions in order to teach social skills to autistic children. This therapeutic origin of the data is a double-edged sword, suggesting both a noble original mission and a danger of improper medicalization once it has been adopted beyond the therapeutic setting.

wants, and simulate concern, respect, friendship, and love.[9] The robot is also exceptionally intelligent, finding an obscure but vital academic reference that upends one of Emma's research programs. Emma occasionally enjoys the attention and expertise that Tom provides and tries to reciprocate. But she ultimately realizes that what Tom is offering is programmed, not a free choice, and is thus fundamentally different than the risk and reward inherent in true human companionship and love.

Emma realizes that, even if no one else knew Tom's nature, her ongoing engagement with it would be dangerous on not only an affective, but also on an epistemic level.[10] As Charles Taylor (1985b, 49) has explained, "experiencing a given emotion involves experiencing our situation as bearing a certain import, where for the ascription of the import it is not sufficient just that I feel this way, but rather the import gives grounds or basis for the feeling."[11] Simply feeling a need for affirmation is not a solid ground or basis for someone else to express affirming emotions. Barring extreme situations of emotional fragility, the other needs to be able to independently decide whether to affirm oneself for that affirmation to have meaning. If simulated expression of such emotions by a thing is done, as is likely, to advance the commercial interest of the thing's owner, there is no solid basis for feeling affirmed either. We can all go from the "wooed" to the "waste" (in Joseph Turow's memorable phrasing) of a firm in the flash of business model shift. Of course, we can also imagine a world in which "haphazardly attached" persons find some solace in the words emitted by LLMs, whatever their nature.[12] But the way such technology fits or functions in such a scenario is far more an indictment (and, ironically, stabilization) of its alienating

[9] I endorse the use of scare quotes (here, for the word "converse") to mark actions taken by robots or AI that would be described without such quotation marks if undertaken by a human. The most accurate approach would be to more fully explain the mechanism and optimization functions of the relevant AI (Tucker 2022); here, for example, to describe Tom's statements as the product of a next-word-prediction algorithm designed to stimulate certain emotional responses from and interaction with Emma. However, given the pressure to describe scenarios expeditiously, and to convey the confusion that is already common in responses to them, the expedient of scare quoting robotic and AI simulations of human action is taken here.

[10] The pronoun "it" is important here, forestalling the improper anthropomorphization that a pronoun like "he" or "him" would encourage. Unfortunately, many persons are already referring to digital personal assistants with personifying pronouns; as one study noted, "Only Google Assistant, having a non-human name, is referred to as *it* by a majority of users. However, users still refer to it using gendered pronouns just under half of the time" (Abercrombie et al. 2021, 27). This is unfortunate because such anthropomorphization can be profoundly misleading regarding the nature and capacities of AI (Abercrombie et al. 2023).

[11] Taylor (1985b, 48) also explains that "by import I mean a way in which something can be relevant or of importance to the desires or purposes or aspirations or feelings of a subject; or otherwise put, a property of something whereby it is a matter of non-indifference to a subject." For more on the epistemic status of emotions, see Martha Nussbaum (2001).

[12] For a fuller understanding of the depth of the problem of loneliness, and particularly male loneliness, in the US, see Kathryn Edin et al. (2019); and also Richard V. Reeves (2022) describing the rise in the percentage of men reporting "no close friends" from 3% in 2001 to 15% in 2015.

environment, than testament to its own excellence or value. As Rob Horning has observed, from an economic perspective, large technology firms "must prefer the relative predictability of selling simulations to the uncontrollable chaos of selling social connection. They would prefer that we interact with generated friends in generated worlds, which they can engineer entirely to suit their ends" (Horning 2024).

While many advocates of "artificial friends" based on affective computing claim that they will alleviate alienation, they are more likely to do the opposite: lure the vulnerable away from truly restorative, meaningful, and resonant human relationships, and into a virtual world. As Sherry Turkle has observed:

> [chatbots] haven't lived a human life. They don't have bodies and they don't fear illness and death . . . AI doesn't care in the way humans use the word care, and AI doesn't care about the outcome of the conversation . . . To put it bluntly, if you turn away to make dinner or attempt suicide, it's all the same to them. (quoted in Mineo 2023)[13]

Like the oxymoronic "virtual reality" of *Ready Player One*, the oxymoronic "artificial empathy" of an "AI friend" is a far-from-adequate individual compensation for the alienating social world such computation has helped create.

## 4.2 SELF-PREVENTING PROPHECY

Despite cautionary tales like *Her* and *I'm Your Man*, myriad persons already engage with "virtual boyfriends and girlfriends" (Ding 2023).[14] As reported in 2023 about just one firm providing these services, Replika:

> Millions of people have built relationships with their own personalized instance of Replika's core product, which the company brands as the "AI companion who cares." Each bot begins from a standardized template – free tiers get "friend," while for a $70 premium, it can present as a mentor, a sibling or, its most popular option, a romantic partner. Each uncanny valley-esque chatbot has a personality and appearance that can be customized by its partner-slash-user, like a Sim who talks back. (Bote 2023)

Chastened in its metaversal ambitions, Meta has marketed celebrity chatbots to simulate conversation online. Millions of persons follow and interact with "virtual influencers," who may be little more than a stylish avatar backed by a PR team (Criddle 2023).

[13] MIT Professor Sherry "Turkle has grown increasingly concerned about the effects of applications that offer 'artificial intimacy' and a 'cure for loneliness.' Chatbots promise empathy, but they deliver 'pretend empathy,' she said, because their responses have been generated from the internet and not from a lived experience. Instead, they are impairing our capacity for empathy, the ability to put ourselves in someone else's shoes."

[14] Xiaoice "has leaned into digital humans and avatars. It leads the 'virtual boyfriend and girlfriend' market with 8 million users. As part of this stream, Xiaoice has an 'X Eva' platform which hosts digital clones of Internet celebrities to provide chat and companionship services."

For any persons who believe they are developing relationships with bots, online avatars, or robots, the arguments in Section 4.1 are bitter pills to swallow. The blandishments of affective computing may well reinforce alienation overall, but sufficiently simulate its relief (for any particular individual) to draw the attention and interest of many desperate, lonely, or merely bored persons. The abstractions of theory cannot match the importuning eyes, perfectly calibrated tone of voice, or calculatedly attractive appearance of online avatars and future robots. Yet human powers of imagination can still divert a critical mass of persons away from the approximations of Nozick's "experience machine" dreamed of by too many in technology firms.

Consider the complexities of human–robot interaction envisioned in the hit HBO series *Westworld*. When asked if it sometimes questions the nature of its reality, the robot named Dolores Abernathy states in Season 1, "Some people choose to see the ugliness in this world. The disarray. I choose to see the beauty. To believe there is an order to our days, a purpose." This refrain could describe a typical product launch for affective computing software, with its bright visions of a happier world streamlined with tech that always knows just what to say, just how to open and close your emails, just what emoji to send when you encounter a vexing text. *Westworld* envisions a theme park where calculated passion goes well beyond the world of bits, culminating in simulated (and then real) murders. The promise of the park is an environment where every bright, dark, or lurid fantasy can be simulated by androids almost indistinguishable from humans. It is the *reductio ad absurdum* (or perhaps *proiectio ad astra*) of the affective surround fantasized by Cusk's depressed divorcee, deploying robotics to achieve what text, sound, and image cannot.

By the third season of *Westworld*'s Möbius strip chronology, Dolores breaks out of the park, driven to reveal to humans of the late twenty-first century that their fates are silently guided by a vast, judgemental, and pushy AI. While the last season of the show was an aesthetic mess, its reticulated message – of humans creating a machine to save themselves from future machines – was a philosophical challenge. How much do we need more computing to navigate the forbiddingly opaque and technical scenarios created by computing itself?

For transhumanists, the answer is obvious: human bodies and brains as we know them are just too fragile and fallible, especially when compared with machines. "Wetware" transhumanists envision a future of infinite replacement organs for failing bodies, and brains jacked into the internet's infinite vistas of information. "Hardware" transhumanism wants to skip the body altogether and simply upload the mind into computers. AIs and robots will, they assume, enjoy indefinite supplies of replacement parts and backup memory chips. Imagine Dolores, embodied in endless robot guises, "enminded" in chips as eternal as stars.[15]

[15] This and the next several paragraphs are drawn from my *Commonweal* article "Is AI Poised to Replace Humanity?" (Pasquale 2023).

The varied and overlapping efficiencies that advanced computation now offer make it difficult to reject this transhumanist challenge out of hand. A law firm cannot ignore large language models and the chatbots based on them, because these tools may not only automate simple administrative tasks now but also may become a powerful research tool in the future. Militaries feel pressed to invest in AI because technology vendors warn it could upend current balances of power, even though the great power conflicts of the 2020s seem far more driven by basic industrial capacities. Even tech critics have Substacks, Twitter accounts, and Facebook pages, and they are all subject to the algorithms that help determine whether they have one, a hundred, or a million readers. In each case, persons with little choice but to use AI systems are donating more and more data to advance the effectiveness of AI, thus constraining their future options even more. "Mandatory adoption" is a familiar dynamic: it was much easier to forego a flip phone in the 2000s than to avoid carrying a smartphone today. The more data any AI system gathers, the more it becomes a "must-have" in its realm of application.

Is it possible to "say no" to ever-further technological encroachments?[16] For key tech evangelists, the answer appears to be no. Mark Zuckerberg has fantasized about direct mind-to-virtual reality interfaces, and Elon Musk's Neuralink also portends a perpetually online humanity. Musk's verbal incontinence may well be a prototype of a future where every thought triggers AI-driven responses, whether to narcotize or to educate, to titillate or to engage. When integrated into performance-enhancing tools, such developments also spark a competitive logic of self-optimization. A person who could "think" their strategies directly into a computing environment would have an important advantage over those who had to speak or type them. If biological limits get in the way of maximizing key performance indicators, transhumanism urges us toward escaping the body altogether.

This computationalist eschatology provokes a gnawing insecurity: that no human mind can come close to mastering the range of knowledge that even a second-rate search engine indexes, and simple chatbots can now summarize, thanks to AI. Empowered with foundation models (which can generate code, art, speech, and more), chatbots and robots seem poised to topple humans from their heights of self-regard. Given Microsoft's massive investments in OpenAI, we might call this a Great Chain of Bing: a new hierarchy placing the computer over the coder, and the coder over the rest of humans, at the commanding heights of political, economic, and social organization.[17]

Speculating about the long-term future of humanity, OpenAI's Sam Altman (2017) once blogged about a merger of humans and machines, perhaps as a way

[16] For an affirmative response in another sociotechnical realm, see Pasquale (2010).

[17] This hierarchy is expertly analyzed by Jenna Burrell and Marion Fourcade (2021) and is closely related to the problem of economics' displacement of other forms of knowledge in policy making. See Marion Fourcade et al. (2015).

for the former to keep the latter from eliminating them outright. "A popular topic in Silicon Valley is talking about what year humans and machines will merge (or, if not, what year humans will get surpassed by rapidly improving AI or a genetically enhanced species)," he wrote. "Most guesses seem to be between 2025 and 2075." This logic suggests a singularitarian mission to bring on some new stage of "human evolution" in conjunction with, or into, machines. Just as humans have used their intelligence to subdue or displace the vast majority of animals, on this view, machines will become more intelligent than humans and will act accordingly, unless we merge into them.

But is this a story of progress, or one of domination? Interaction between machines and crowds is coordinated by platforms, as MIT economists Erik Brynjolffson and Andrew McAfee have observed. Altman leads one of the most hyped ones. To the extent that CEOs, lawyers, hospital executives, and others assume that they must coordinate their activities by using large language models like the ones behind OpenAI's ChatGPT, they will essentially be handing over information and power to a technology firm to decide on critical future developments in their industries (Altman 2017). A narrative of inevitability about the "merge" serves Altman's commercial interests, as does the tidal wave of AI hype now building on Elon Musk's X, formerly known as Twitter.

The middle-aged novelist who narrates Hari Kunzru's (2020) *Red Pill* wrestles with this spectre of transhumanism, and is ultimately driven mad by it. Suffering writer's block, he travels from his home in Brooklyn to Berlin, for a months-long retreat. Lonely and unproductive at the converted mansion he's staying at, he becomes both horrified and fascinated by a nihilistic drama called *Blue Lives*, which features brutal cops at least as vicious as the criminals they pursue. Its dialogue sprinkled with quotes from Joseph de Maistre and Emil Cioran, *Blue Lives* appears to the narrator as something both darker and deeper than the average police procedural. He gradually becomes obsessed with the show's director, Anton.

Anton is an alt-rightist, fully "red pilled," in the jargon of transgressive conservatism. He also dabbles in sociobiological reflections on the intertwined destiny of humans and robots. The narrator relates how Anton described his views in a public speaking tour:

> [Anton] spoke about his "program of self-optimization." He worked out and took a lot of supplements, but when it came to bodies, he was platform-agnostic. Whatever the substrate, carbon-based or not, he thought the future belonged to those who could separate themselves out from the herd, intelligence-wise … Everything important would be done by a small cognitive elite of humans and AIs, working together to self-optimize. If you weren't part of that, even selling your organs wasn't going to bring in much income, because by then it would be possible to grow clean organs from scratch. (207)

In a narcissistic short film celebrating himself, Anton announces that: "Around us, capital is assembling itself as intelligence. That thought gives me energy. I'm growing stronger by the day" (206).

The brutal logic here is obvious: some will be in charge of the machines, perhaps merging with them; most will be ordered around by the resulting techno-junta.[18] Dismissing "unproductive" humans as so many bodies is the height of cruelty (207). But it also fits uncomfortably well with a behaviourist robot rights ideology that claims that only what an entity *does* is what matters, not what it *is* (the philosophical foundation of Anton's "platform agnosticism"). Nick Cave elegantly refutes this behaviourism in an interview exploring his recent work:

> Maybe A.I. can make a song that's indistinguishable from what I can do. Maybe even a better song. But, to me, that doesn't matter – that's not what art is. Art has to do with our limitations, our frailties, and our faults as human beings. It's the distance we can travel away from our own frailties. That's what is so awesome about art: that we deeply flawed creatures can sometimes do extraordinary things. A.I. just doesn't have any of that stuff going on. Ultimately, it has no limitations, so therefore can't inhabit the true transcendent artistic experience. It has nothing to transcend! It feels like such a mockery of what it is to be human. (Petrusich and Cave 2023)[19]

As Leon R. Kass (2008) articulates, "Like the downward pull of gravity without which the dancer cannot dance, the downward pull of bodily necessity and fate makes possible the dignified journey of a truly human life." For "make a song" in Cave's passage, we could include so many other human activities: run a mile, play a game of chess, teach a class, console a mourning person, order a drink. We are so much more than what we do and make, bearing value that Anton appears unable or unwilling to recognize.

Alarmed by the repugnance of Anton's message, the narrator becomes distressed by his success. He argues with him at first, accusing him of trying to "soften up" his *Blue Lives* audience to accept a world where "most of us [are] fighting for scraps in an arena owned and operated by what you call a 'cognitive elite'." (Kunzru 2020, 208). He calls out Anton's fusion of hierarchical conservatism and singularitarianism as a new Social Darwinism. But he cannot find a vehicle to bring his own counter-message to the world. The accelerationist logic of vicious competition, first among humans, then among humans enhanced by machines, and finally by machines themselves,

[18] For a critical perspective on this logic of AI, rooted in a Marxian account of automation, see Matteo Pasquinelli (2023).

[19] See also David Means (2023): "A.I. will never feel the sense of mortality that forms around an unfinished draft, the illogic and contradictions of the human condition, and the cosmic unification of pain and joy that fuels the artistic impulse to keep working on a piece until it is finished and uniquely my own."

signalling the obsolescence of the human form, is just too strong for him.[20] By the end of the novel, his attempt at a *cri de coeur* crumples into capitulation:

> With metrication has come a creeping loss of aura, the end of the illusion of exceptionality which is the remnant of the religious belief that we stand partly outside or above the world, that we are endowed with a special essence and deserve recognition or protection because of it. We will carry on trying to make a case for ourselves, for our own specialness, but we will find that arrayed against us is an inexorable and inhuman power, manic and all-devouring, a power thirsty for the total annihilation of its object, that object being the earth and everything on it, all that exists. (Kunzru 2020, 227)

The intertwined logic of singularitarianism, DeMaistrean conservatism, and contempt for humanity, seem to him inescapable. But Kunzru has his narrator come to this "realization" just as he is slipping into madness.

There are some visions of the future one must simply reject and cannot really argue with; their premises are simply too far outside the bounds of moral probity.[21] Eugenicist promotion of a humanity split by its degree of access to technology is among such visions. It is a dystopia (as depicted in series like *Cyberpunk: Edgerunners* and films like *Elysium*), not a rational policy proposal. The task of the intellectual is not to toy with such secular eschatologies, calculating the least painful glidepath toward them, or amelioration of their worst effects, but to refute and resist them to prevent their realization. The same can be said of "longtermist" rationales for depriving current disadvantaged persons' of resources in the name of the eventual construction of trillions of virtual entities (Torres 2021, 2022). Considering them too deeply, for too long, means entertaining a devaluation of the urgent needs of humanity today – and thus of humanity itself.

## 4.3 CONCLUSION

It will take a deep understanding of political economy, ethics, and psychology (and their mutual influence) to bound our emotional engagement with ever more personalized and persuasive technology. In an era of alexithymia, machines will increasingly promise to name and act upon our mental states.[22] Broad awareness of

[20] For a fuller articulation (and critique) of this accelerationist vision of future evolution, see Benjamin Noys (2014).

[21] As Charles Taylor (1985b) observes, in the social sciences "in so far as they are hermeneutical there can be a valid response to 'I don't understand' which takes the form, not only 'develop your intuitions,' but more radically 'change yourself.' This puts an end to any aspiration to a value-free or 'ideology-free' social science" (54).

[22] For a compelling description of the political entailments of alexithymia, see Manos Tsakiris (2020): "The psychological concept of alexithymia (meaning 'no words for feelings') captures this difficulty in identifying, separating or verbally describing our feelings. An emotional prescription (such as 'you should feel . . .') and affect-labelling (such as 'angry') can function

the machines' owners' agendas will help prevent a resulting colonization of the lifeworld by technocapital (Pasquale 2020a). Culture can help inculcate that awareness, as the films and novels discussed have shown.[23]

The chief challenge now is to maintain critical distinctions between the artificial and the natural, the mechanical and the human. One foundation of computational thinking is "reformulating a seemingly difficult problem into one we know how to solve, perhaps by reduction, embedding, transformation, or simulation" (Wing 2004, 33). Yet there are fundamental human capacities that resist such manipulation, and particularly put us on guard against simulation. Reduction of an emotional state to, say, one of six "reaction buttons" on Facebook often leaves out much critical context.[24] Simulation of care by a robot does not amount to care, because it is not freely chosen. Carissa Veliz's (2023) suggestion that chatbots not use emojis is wise because it helps expose the deception inherent in representation of non-existent emotional states.

To be obliged to listen to robots as if they were persons or to care about their "welfare," is to be distracted from more worthy ends and more apt ways of attending to the built environment. Emotional attachments to AI and robotics are not merely dyadic, encapsulated in a person's and a machine's interactions. Rather, they reflect a social milieu, where friendships may be robust or fragile, work/life balance well-respected or non-existent, conversations with persons free-flowing or clipped. It should be easy enough to imagine in which of those worlds robots marketed as "friends" or "lovers" would appear as plausible as human friends and lovers. That says more about their nature than whatever psychic compensations they afford.

## REFERENCES

Abercrombie, Gavin, Amanda Cercas Curry, Tanyi Dinkar, Verena Rieser, and Zeerak Zakat. "Mirages: On Anthropomorphism in Dialogue Systems." *Arxiv* (2023). https://arxiv.org/abs/2305.09800.

Abercrombie, Gavin, Amanda Cercas Curry, Mugdha Pandya, and Verena Rieser. "Alexa, Google, Siri: What Are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants." In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, edited by Marta Costa-jussà, Hila Gonen, Christian Hardmeier, and Kellie Webster, 24–33. Online: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.gebnlp-1.4.

as the context within which people will construct their emotions, especially when we're interoceptively dysregulated."

[23] Critics of my approach may question the epistemic status of narratives in developing moral intuitions and policy positions. While space limitations preclude a full response here, I have made a case for the relevance of literature to moral and policy inquiry in Pasquale (2020b).

[24] For just one of many examples of the type of context that may matter, see Jerome Kagan (2019). For powerful critiques of reductionism in many affective computing scenarios, see Andrew McStay (2023).

Altman, Sam. "The Merge." *Sam Altman* (blog), July 12, 2017. https://blog.samaltman.com/the-merge.

Birhane, Abeba, Jelle van Dijk, and Frank Pasquale. "Debunking Robot Rights Metaphysically, Ethically, and Legally." *First Monday* 29, no. 4 (2024). https://doi.org/10.5210/fm.v29i4.13628.

Bote, Joshua. "Replika Wanted to End Loneliness with a Lurid AI Bot. Then Its Users Revolted." *San Francisco Gate*, April 27, 2023. www.sfgate.com/tech/article/replika-san-francisco-ai-chatbot-17915543.php.

Burk, Dan L. "Asemic Defamation, or, the Death of the AI Speaker." *First Amendment Law Review* 22 (2025): 189–232.

Burrell, Jenna, and Marion Fourcade. "The Society of Algorithms." *Annual Review of Sociology* 47 (2021): 213–237. https://doi.org/10.1146/annurev-soc-090820-020800.

Christian, Brian. "How a Google Employee Fell for the Eliza Effect." *The Atlantic*, June 21, 2022. www.theatlantic.com/ideas/archive/2022/06/google-lamda-chatbot-sentient-ai/661322/.

Criddle, Cristina. "How AI-Created Fakes Are Taking Business from Online Influencers." *Financial Times*, December 29, 2023. www.ft.com/content/e1f83331-ac65-4395-a542-651b7df0d454.

Cusk, Rachel. *Transit*. New York: Farrar Strauss Giroux, 2017.

Danaher, John. "Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism." *Science and Engineering Ethics* 26 (2020): 2023–2049. https://doi.org/10.1007/s11948-019-00119-x.

Danaher, John, and Neil Macarthur, eds. *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press, 2017.

Ding, Jeffrey. "XiaoIce, Where Do We Go from Here?" *ChinAI* (blog), December 18, 2023. https://chinai.substack.com/p/chinai-248-xiaoice-where-do-we-go.

Edin, Kathryn, Timothy Nelson, Andrew Cherlin, and Robert Francis. "The Tenuous Attachments of Working-Class Men." *Journal of Economic Perspectives* 33, no. 2 (2019): 211–228.

Fourcade, Marion, Etienne Ollion, and Yann Algan. "The Superiority of Economists." *Journal of Economic Perspectives* 29, no. 1 (February 2015): 89–114. https://doi.org/10.1257/jep.29.1.89.

Gebru, Timnit, and Émile P. Torres. "The TESCREAL Bundle: Eugenics and the Promise of Utopia through Artificial General Intelligence." *First Monday* 29, no. 4 (2024). https://doi.org/10.5210/fm.v29i4.13636.

Harris, Jamie, and Jacy Reese Anthis. "The Moral Consideration of Artificial Entities: A Literature Review." *Science and Engineering Ethics* 27, no. 53 (2021).

Horning, Rob. "The Dialectic of Simulation." *Internal Exile*, June 19, 2024. https://robhorning.substack.com/p/dialectic-of-simulation.

Ishiguro, Kazuo. *Klara and the Sun*. New York: Knopf, 2021.

Jankowicz, Nina. *How to Lose the Information War: Russia, Fake News, and the Future of Conflict*. London: I. B. Tauris, 2020.

Kagan, Jerome. *Kinds Come First: Age, Gender, Class, and Ethnicity Give Meaning to Measures*. Cambridge, MA: MIT Press, 2019.

Kass, Leon R. "Defending Human Dignity." In *Human Dignity and Bioethics: Essays Commissioned by the President's Council on Bioethics*, edited by President's Council on Bioethics. U.S. Government Printing Office, 2008.

Kirschenbaum, Matthew. "Prepare for the Textpocalypse." *The Atlantic*, March 2023. www.theatlantic.com/technology/archive/2023/03/ai-chatgpt-writing-language-models/673318/.

Kunzru, Hari. *Red Pill*. London: Scribner, 2020.

Levy, David. *Love and Sex with Robots*. New York: Harper Perennial, 2008.
Lyotard, Jean-François. *The Inhuman: Reflections on Time*, translated by Geoffrey Bennington and Rachel Bowlby. Redwood City, CA: Stanford University Press, 1992.
Manguso, Sarah. *300 Arguments*. New York: Picador, 2018.
McStay, Andrew. *Automating Empathy*. New York: Oxford University Press, 2023.
Means, David. "A.I. Can't Write My Cat Story Because It Hasn't Felt What I Feel." *N. Y. Times*, March 26, 2023. www.nytimes.com/2023/03/26/opinion/ai-art-fiction.html.
Mineo, Liz. "Why Virtual Isn't Actual, Especially When It Comes to Friends." *Harvard Gazette*, June 21, 2023. https://news.harvard.edu/gazette/story/2023/12/why-virtual-isnt-actual-especially-when-it-comes-to-friends/.
Noys, Benjamin. *Malign Velocities: Accelerationism and Capitalism*. Winchester: Zero Books, 2014.
Nussbaum, Martha. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press, 2001.
Parsons, William B. "The Oceanic Feeling Revisited." *Journal of Religion* 78, no. 4 (1998): 501–523.
Pasquale, Frank. "Is AI Poised to Replace Humanity?" *Commonweal*, November 22, 2023. www.commonwealmagazine.org/ai-poised-replace-humanity.
"The Algorithmic Self." *Hedgehog Review* 17, no. 1 (2015).
"The Automated Public Sphere." In *The Politics of Big Data*, edited by Ann Rudinow Sætnan, Ingrid Schneider, and Nicola Green, 19–46. London: Taylor & Francis, 2018.
"Cognition-Enhancing Drugs: Can We Say No?" *Bulletin of Science, Technology & Society* 30, no. 9 (2010): 9–13. https://doi.org/10.1177/0270467609358113.
*New Laws of Robotics: Defending Human Expertise in the Age of AI*. Cambridge, MA: Belknap Press, 2020a.
"The Substance of Poetic Procedure: Law & Humanity in the Work of Lawrence Joseph." *Law and Literature* 32, no. 1 (2020b): 1–46. https://doi.org/10.1080/1535685X.2019.1680130.
"Two Concepts of Immortality." *Yale Journal of Law & the Humanities* 14, no. 1 (2002): 73–121.
Pasquinelli, Matteo. *The Eye of the Master: A Social History of Artificial Intelligence*. New York: Verso, 2023.
Petrusich, Amanda, and Nick Cave. "Nick Cave on the Fragility of Life." *New Yorker*, 23 March 2023. www.newyorker.com/culture/the-new-yorker-interview/nick-cave-on-the-fragility-of-life.
Pugh, Allison. *The Last Human Job: The Work of Connecting in a Disconnected World*. Princeton, NJ: Princeton University Press, 2024.
Reeves, Richard V. *Of Boys and Men: Why the Modern Male Is Struggling, Why It Matters, and What to Do about It*. Washington, DC: Brookings Institution Press, 2022.
Rogers, Reece. "How to Detect AI-Generated Text, According to Researchers." *Wired*, 2022. www.wired.com/story/how-to-spot-generative-ai-text-chatgpt/.
Singer, Peter W. *LikeWar: The Weaponization of Social Media*. Boston: Houghton Mifflin Harcourt, 2018.
Tangermann, Victor. "Transcript of Conversation with "Sentient" AI Was Heavily Edited." *Futurism*, June 14, 2022. https://futurism.com/transcript-sentient-ai-edited.
Taplin, Jonathan. *The End of Reality: How Four Billionaires Are Selling a Fantasy Future of the Metaverse, Mars, and Crypto*. New York: PublicAffairs, 2023.
Taylor, Charles. *Philosophy and the Human Sciences*. Cambridge: Cambridge University Press, 1985a.
"Self-Interpreting Animals." In *Human Agency and Language: Philosophical Papers Vol. 1*, 45–76. Cambridge: Cambridge University Press, 1985b.

Tian, Edward. "GPTZero Case Study: Models and Exploits." *GPTZero* (blog), February 7, 2023. https://gptzero.substack.com/p/gptzero-case-study-models-and-exploits.

Torres, Émile P. "The Acronym behind Our Wildest AI Dreams and Nightmares." *Truthdig*, June 15, 2023. www.truthdig.com/articles/the-acronym-behind-our-wildest-ai-dreams-and-nightmares/.

"Against Longtermism." *Aeon*, 2021. https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo.

"Understanding 'Longtermism:' Why This Suddenly Influential Philosophy Is So Toxic." *Salon*, August 20, 2022. www.salon.com/2022/08/20/understanding-longtermism-why-this-suddenly-influential-philosophy-is-so/.

Tsakiris, Manos. "Politics Is Visceral." *Aeon*, September 2020. https://aeon.co/essays/politics-is-in-peril-if-it-ignores-how-humans-regulate-the-body.

Tucker, Emily. "Artifice and Intelligence." *Tech Policy Press*, March 16, 2022. www.techpolicy.press/artifice-and-intelligence/.

Veliz, Carissa. "Chatbots Shouldn't Use Emojis: Artificial Intelligence That Can Manipulate Our Emotions Is a Scandal Waiting to Happen." *Nature*, March 14, 2023. www.nature.com/articles/d41586-023-00758-y.

White, James Boyd. "What Can a Lawyer Learn from Literature?" *Harvard Law Review* 102, no. 8 (1989): 2014–2047.

Wing, Jeannette M. "Computational Thinking." *Communications of the ACM* 49, no. 3 (2004): 33–35.