DOI: 10.1215/2834703X-10734016

# Editor's Introduction: Humanities in the Loop

LAUREN M. E. GOODLAD

*If it should turn out to be true that knowledge (in the modern sense of know-how) and thought have parted company for good, then we would indeed become helpless slaves, not so much of our machines as of our know-how, thoughtless creatures at the mercy of every gadget which is technically possible, no matter how murderous it is.*
—Hannah Arendt, *The Human Condition*

*Artificial intelligence (AI) is an emerging set of computer technologies that affect individuals, communities, and societies at a global scale. Touted as a fourth industrial revolution, AI is nonetheless poorly understood and subject to hype, clickbait, and anxiety. Although there is increasing talk of making AI "ethical," "democratic," "human-centered," and "responsible," these conversations suffer from a lack of cross-disciplinary dialogue, critical understanding, and public engagement.*
—Excerpt from the proposal for a new interdisciplinary journal to be named *Critical AI*

*AI is the tech the world has always wanted.*
—Sam Altman, CEO, OpenAI



For the last decade or so, talk about artificial intelligence (AI) has surged, powered by a number of "machine learning" (ML) breakthroughs made possible by advances in computing and the accumulation of vast stores of data generated on the internet. The forecast of an AI-driven "revolution" was already making waves before the November 2022 release of ChatGPT unleashed a tsunami of hype around large language models (LLMs), a type of data-driven ML technology. That hype cycle remains in full force as we prepare *Critical AI*'s first issue, "Data Worlds," for publication in August 2023.[1] For a specific introduction to "Data Worlds," see the article I coauthored with Katherine Bode. The goal of this editor's introduction is to welcome readers to a new interdisciplinary undertaking. *Critical AI*'s contributors, editorial team, and reviewers have worked hard to make these virtual pages legible and relevant to your work and your world. But you, our readers, will be the ones to judge.

If you, reader, are a humanist, critical theorist, or interpretive social scientist, please read on (I will address you directly toward the end of this introduction). If you are a technologist, you may wonder what specifically is meant by my title, "Humanities in the Loop." You will of course recognize the pun on "human in the loop," a common feature in the design of automated systems.[2] But you may question the implied corollary—that close interaction with humanities thinking can enrich your research or critical perspectives. If so, I hope "Data Worlds" rewards your curiosity. In the introduction, coeditor Katherine Bode and I sketch out a broad vision of *critical AI studies*, an emerging body of interdisciplinary practices that we invite you to help develop and disseminate.

The community of practice that *Critical AI* thus addresses hopes to bring critical thinking of the kind that interpretive disciplines foster into dialogue with work by technologists and others, including community organizers, educators, entrepreneurs, health professionals, journalists, lawyers, legislators, and indeed anyone who, as the journal's mission statement puts it, "shares the understanding of interdisciplinary research as a powerful tool for building and implementing accountable technology in the public interest."[3] More ambitiously, the shared aspiration behind *Critical AI* is to shape and activate conversations—in academia, industry, policy-making, media, and the public at large—at a time when the rapid commercialization and deployment of "AI" products coincide with, and sometimes abet, environmental crisis, economic precarity, political authoritarianism, intensifying harms, and an already disconcerting concentration of wealth and power.

To be clear, *Critical AI* does not invite Luddism (we do not seek wholesale rejection of research in ML or other autonomous systems) or, still less, doomerism (we do not believe AI is likely to pose

existential threats to the human species or even imminent threats to most forms of human labor).[4] Attentive to history, we know that accountable technologies developed in the public interest have played a key role in human, more-than-human, and environmental flourishing and believe that versions of what is now called AI can do so as well.[5] We nonetheless perceive the significant need to question and counter boosterism: unaccountable, technodeterministic, and disempowering discourse and hype (including doomerist hype) that undermine informed public discussion, knowledgeable journalism, and democratic oversight. We further recognize that automated systems *already* impose unacceptable harms on individuals, public infrastructures, and environmental well-being.[6] And we perceive that the "democratization" that AI marketing glibly extols can damage real democratic processes, ignore affected communities, and circumvent the social, political, and ecological interests at stake. Without articulating any simple or single agenda, our goal is to publish diverse forms of peer-reviewed research that represent the best possible interdisciplinary critical thinking on "AI" and its contexts broadly conceived.

The occasional use of quotation marks around "artificial intelligence" and "AI" may puzzle some readers. Why title a journal *Critical AI* and then advise caution with respect to the pivotal term? The reason is complicated, but I will try to explain. The term "AI" was coined by the organizers of a Dartmouth College "summer research project" on the topic (McCarthy et al. 2006). Despite the limited success of these early endeavors, "AI" entered the popular imagination in large part because of the (already existing) public delight in stories about intelligent "robots," sentient "androids," and superintelligent computers.[7] Thus, both the popular understanding of AI and its technological horizon are deeply conditioned by an influential body of fictional works in which Promethean science quests after the secrets of life itself—usually to dystopian effect. Nevertheless, while science fiction has shaped the popular imaginary of "AI" for decades, no fictional work (so far) has directly engaged the data-driven statistical modeling that claims the title of "AI" in the present day. The pronounced disconnect between, on the one hand, iconic storytelling (whether the amiable robots of *Star Wars* or the dolorous androids of *Blade Runner* and *Westworld*), and on the other hand, actually existing technologies (such as COMPAS, AlphaZero, or GPT-4) sets the stage for disempowering confusion.[8]

The story of AI's inception often begins in 1943, when Alan M. Turing's idea (1936) for the mathematical encoding of instructions and data to activate electronic circuits and switches helped to inspire the first mathematical model of a *neural network* (McCulloch and Pitts 1943; cf. Davis 2000), a term that nods to the fundamental unit of the human brain. Notably, Turing's own conception of the human/machine relation is more complex. When his famous essay in *Mind* (Turing 1950) takes up the question of whether machines can "think," Turing pivots to the very different topic of whether individuals unknowingly engaged in a question-and-answer exchange with a text-generating program can be persuaded that they are conversing with a human being. Though this idea became the basis for a "Turing test" for human-like or "general" artificial intelligence, Turing himself never spoke of a test and described his thought experiment as an "imitation game." Thus, while the organizers of the Dartmouth workshop envisioned "AI" as a means "to make machines use language, form abstractions and concepts," and to solve "problems now reserved for humans," Turing (who died in 1954) appears to have been interested in programming machines to *imitate* and *perform* human-like conversation—not (or at least not explicitly) to reproduce or enact human intelligence.

Imitating conversation was the obvious goal of ELIZA (Weizenbaum 1976: 3), an interactive software program designed by MIT computer scientist Joseph Weizenbaum in 1966 to "parody" the speech of a "Rogerian psychotherapist." To his chagrin, Weizenbaum discovered that "extremely short exposures" to this simple conversational program "could induce powerful delusional thinking in quite normal people" (7). The tendency for humans to anthropomorphize and project understanding onto systems that generate dialogue in natural language came to be known as the "ELIZA effect" (Hofstadter 1996). But for Weizenbaum, whose *Computer Power and Human Reason: From Judgment to Calculation* (1976) remains a foundational work in AI ethics, the problem went deeper. Building on the midcentury thought of Hannah Arendt and Lewis Mumford, Weizenbaum argued that AI research had become symptomatic of a fixation with technology that was producing "ever more highly rationalistic" and "mechanistic" views of people and the world (11). In doing so, he

rejected the analogy between calculation and human reason (and thus between computer and brain) that had become central to research in the field. The subtitle for Weizenbaum's book, *From Judgment to Calculation*, adapts Arendt's distinction from a 1971 critique of the military decision-making that escalated the Vietnam War. But Weizenbaum's case against quantitative reductionism also echoes Arendt's *The Origins of Totalitarianism* (1951), in which she describes Europe's catastrophic encounters with imperial will to power and dehumanizing scientistic propaganda.[9]

The 1970s also gave rise to an influential British report (Lighthill 1973) that criticized the "grandiose aims" and "inflated predictions" of AI research and is often adduced as the trigger for the first of two AI "winters." The loss of funding and interest that characterized these periods—one in the mid-1970s, and the second in the mid-1980s—produced a chastened language in which the technically specific vocabulary of "machine learning" replaced an "AI" discourse that had become too ambitious and mixed up with science fiction to provide a suitable term for research. The machine "learning" in question (in contrast to the generalizable experiential learning of humans) involves programmed instructions to update statistical weights in order to optimize predictions about specific data-driven tasks or domains.

In fact, "learning" in this technical sense continues to be the mainstay of AI's most impressive feats, whether the predictions in question involve playing Go, mapping the geography of a novel protein, identifying cancer, or generating plausible textual sequences in response to a prompt. As Nicolas Malevé and Katrina Sluis describe in their article in this issue, the year 2012 became a landmark in the rise of "deep learning"—an ML approach involving multilayered ("deep") software architectures. Beginning in the 2000s (as Bode and I discuss in our introduction), the increasing availability of data from the internet—alongside major improvements in hardware and computing power—enabled artificial "neural networks" to bear fruit seventy years after their initial conception. This confluence of factors, as Meredith Whittaker (2021: 51–52) puts it, showed "the commercial potential" of DL along with "the power of AI" as a "marketing hook."

More than a decade into the DL paradigm's romance with massive scale, the world's largest tech companies are using the rhetoric of "artificial intelligence," and even "artificial general intelligence" (AGI), to market disembodied, virtual, and affectless statistical models, trained on vast datasets, to a public accustomed to associating such terms with fictional androids. Given that "AGI" has historically denoted human-level modes of intelligence that remain well beyond the functionalities of data-driven systems,[10] the current escalation of marketing rhetoric exacerbates three core dilemmas for critical AI studies and *Critical AI* to keep in view.

**Reductive and Controversial Meanings of "Intelligence"**

While the "intelligence" at stake in "artificial intelligence" has never been rigorously theorized or defined, the recent harnessing of "AI" to commercial interests has intensified the field's immersion in hype. To be sure, it is possible to speak scientifically about machine "intelligence."[11] Nonetheless, the present-day rhetoric of "AI" research has for decades been shaped by loose anthropomorphisms that have little basis in fact. For example, the idea that artificial neural networks (Sejnowski 2018: ix, 3) "reverse engineer" the human brain and, therefore, learn the way "babies" do draws on the underlying assumption that brains (and human intelligence more generally) are analogous to computers and vice versa. The analogy was inspired by McCulloch and Pitts' 1943 model of the artificial neuron; taken up in influential works by Norbert Wiener (1948) and John von Neumann (1958); and reiterated in the research of cognitive and some computer scientists committed to deep learning.[12] Thus, despite robust interventions of various kinds, including nuanced distinctions between calculation and judgment (Weizenbaum 1976; Smith 2019); technical discussions of the limits of data-driven pattern-finding (Pearl and Mackenzie 2018; Marcus and Davis 2019; Lanier 2023); cogent critiques of machine "understanding" (Bender and Koller 2020); biting commentary on Silicon Valley's fixation with "superintelligence" (Chiang 2017); and exhaustive accounts of the dangers of LLMs (Bender et al. 2021; Weidinger et al. 2022), the research and commercialization of AI continues to adhere to reductive assumptions such as John McCarthy's (1997) definition of *intelligence* as "the computational part of the ability to achieve goals in the world."[13]

AI discourse thus remains stubbornly rooted in a simplistic anthropocentric mindset that regards human intelligence as the very paradigm of *any* intelligence (ignoring the diverse intelligence of, for example, nonhuman animals). The same mindset conceives of human intelligence as a fundamentally calculative and measurable capacity located in the brain—an assumption that encourages researchers to mobilize the atomized abstractions of narrow utilitarianism, economics, and game theory, while ignoring more complex and situated perspectives (including the impact of physical embodiment, emotions, webs of relationality, relations of care, cultural contexts, and/or philosophical assumptions).[14] In its reductive form, "intelligence" harks back to pseudoscientific hierarchies and norms derived from the long histories of biometrics, eugenics, imperialism, and their totalitarian outcomes.[15] By contrast, critical perspectives recognize the plurality and contextualism of intelligence, human and otherwise.[16] With respect to "AI," critical perspectives perceive how anthropomorphic analogies misrepresent the functionalities of data-driven machine systems when they conflate predictive analytics with human decision-making and equate massive datasets with human knowledge, social experience, and cultural commitments. The point of rejecting such flawed assumptions is as much to capture robust understandings of machine intelligence as it is to biological and sociocultural lives.

complicate mechanistic simplifications

**Problematic Benchmarks and Tests for Supposedly Scientific Terms Such as "AGI"**

In an important essay, Raji et al. (2021) show how the benchmarks used to assess progress toward the "general" intelligence associated with humans lack *construct validity*, that is, the ability to measure what is claimed. As datasets that represent a selective slice of the world and tasks that simplify real-world complexity are used to suggest that human-like "AI" is near at hand, the field's embrace of these flawed standards of progress discourages rigor, foments hype, and (as Weizenbaum predicted fifty years ago) promotes mechanistic understandings of people and the world. In such a climate, Turing's "imitation game" is put forward as a serious "test" of human-like intelligence, despite the well-known propensity of people (dating back to ELIZA) to project human consciousness onto machines that generate even minimally sensible language. Popular media exacerbate these weaknesses, as when journalists imply that a language model's ability to pass examinations in, say, law or medicine warrants professional competence (or even reliable output of codified knowledge).

When Marcus and Davis (2019: 3) define "AGI" as "general-purpose artificial intelligence with the flexibility of human intelligence," the generality in question equates with the "strong" intelligence that Pearl and Mackenzie (2018: 30) associate with the ability to transpose knowledge and reason across particular domains, experiences, and activities. Since the 1990s, "AGI" has thus stood for a human-like standard far beyond the ability to incrementally expand "artificial narrow intelligence" (ANI)—that is, purpose-built systems that excel at one task. Rather, "AGI" has typically signified a kind of holy grail in which computerized systems simulate the human capacity to ponder causes, undertake thought experiments, and successfully navigate and reason about the world's myriad objects. With the advent of powerful language models, however, the valence of "generality" or "general intelligence" has become ever more murky and contested. Because state-of-the-art LLMs can be used to generate text, answer questions, generate code, translate language, and perform some low-level mathematical operations, this advance from single-purpose to multipurpose systems is now misleadingly conflated with "general" intelligence in the human sense. However, an LLM's ability to perform multiple tasks for which it has trained on relevant data does not imply that such a system demonstrates (or soon will demonstrate) a human-like ability to transpose knowledge and reasoning cultivated in one domain of experience to an entirely novel task. Still less does it imply an emergent capacity to undertake metacognitive reflection.

Even so, recent papers—some authored by employees of Google, Microsoft, and Open AI—have exploited weak benchmarks to claim the immanence of "AGI."[17] At the same time, some technology companies are altering the meaning of "AGI" entirely. According to OpenAI's CEO (Altman 2023), "AGI" refers to "AI systems that are generally smarter than humans"—a vague standard that could describe almost any state-of-the-art model trained on large datasets. When such ambiguous perspectives claim to benefit humanity, AGI discourse reinforces a technodeterministic worldview in

which projects of futuristic social engineering work hand in hand with neo-eugenic theories and pseudosciences (Gebru and Torres 2023; Torres 2023). As we have seen, hyped claims for superintelligent AGI often fuse with doom-laden anxieties over AI's supposedly "existential" risk to the human species. AI hype today is thus remarkably Janus-faced: steeped in utopian and dystopian idioms as well as peculiar blends of the two. Nonetheless, what such discourses share is the tendency to minimize the real-world harms of actually existing technologies—a topic that leads us to a third major concern.

**Bias, Errors, and Concentration of Power**

The tendency of AI champions such as Sam Altman and Elon Musk to fuse boosterism and doomerism creates a media ecosystem in which real-world harms to people and the planet must vie for attention with clickbait-friendly speculations about the supposedly existential risks of nonexistent technologies. By contrast, scholars including Costanza-Chock, Raji, and Buolamwini (2022: 1572) have documented how unaccountable decision-making systems "wrongfully deny welfare benefits, kidney transplants, and mortgages to individuals of color as compared to their white counterparts" and trigger "wrongful arrests due to biases in facial recognition technologies"; while Raji et al. (2022: 959–60) have detailed a "functionality problem" such that "scholars, the press, and policymakers" often leap to assessing "ethical" questions about new products without first establishing that the systems in question actually work.[18]

As a host of papers have shown since Bender et al.'s (2021) classic paper on the dangers of LLMs, the widespread practice of scraping data from the internet "bakes in" dubious content with a range of harms that include "persistent toxic" content (Gehman et al. 2020: 3356), "severe" bias against Muslims (Abid, Farooqi, and Zou 2021: 298), and the frequent generation of "misconceptions" (including the mimicking of conspiracy theories, climate change denial, and pseudoscientific beliefs) (Lin, Hilton, and Evans 2022). Hundt et al. (2022: 743, 752) warn that robots programmed with CLIP (an OpenAI image-to-text classifier integral to image-generating systems such as DALL-E) pick up "malignant stereotypes" including "racist, sexist, and scientifically discredited physiognomic behavior," a problem primed to imprint "a permanent blemish on the history of Robotics."

Whereas Hundt et al. (753, 743–44) propose that these flawed systems "be paused, reworked, or even wound down" until "outcomes can be proven safe, effective, and just," tech companies imply that mitigation can manage harms. Yet as journalists begin to unearth the tactics through which much-discussed products like ChatGPT improve performance and lessen toxicity, the public learns that "mitigation" relies heavily on exploitative human labor. For example, *Time* (Perrigo 2023a) reported on the Kenyans earning less than two dollars per hour to label graphic content for OpenAI at high speed (including content depicting "child sexual abuse, bestiality, murder, suicide, torture, self harm, and incest"), while NBC News (Ingram 2023) covered the "hidden army" of US contract workers earning fifteen dollars per hour with no benefits to label data and improve outputs (cf. Irani and Silberman 2013; Gray and Suri 2019). As Malevé and Sluis explain in this issue, the deep learning "revolution" of 2012 was built on industrial-scale crowdwork of this very kind. What is hyped as "AI" and even "AGI" is thus the product not only of technology companies and their investors, but also—and more fundamentally—of the many millions of people subject to copyright infringement and the nonconsensual use of their data, and the low-wage and high-stress modes of "human in the loop" through which systems for probabilistic mimicry improve their performance in an imitation game.[19]

* * *

I conclude with a parting word to another group among *Critical AI*'s potential readers. If you, dear reader, regard yourself as a humanist of some stripe—perhaps a literary critic, historian, political theorist, philosopher, or digital humanist—the invocation of "humanities" discourse may strike you as strangely belated. After all, would not the "humanist" readers of a new interdisciplinary journal recognize themselves and their most cogent ideas as, by now, *post*human in every conceivable way—as fragmented, reassembled, and distributed as many digital processes? As commodified and

datafied as any late-capitalist artifact? As stripped of any pretense to biological or cultural privilege as the barest of bare life?

For a variety of reasons, I think the answer is "no," or at least "not quite." The long histories of "the humanities" and of "the human" are, to be sure, plagued by many problems, most of which have been extensively documented by deconstructionists, feminists, postcolonial theorists, and critical race scholars during the last several decades. But that does not mean we can rectify these problems or, still less, address present crises by ignoring the human-dominated structures of power that enable the status quo. As an interdisciplinary journal, *Critical AI* will not paper over the complicities or blind spots of any theory, method, or critical practice. Nonetheless, the urgent conversations we hope to spark across disciplines and domains can, I think, draw on what is strongest in the humanities and posthumanities.

As Christoper Newfield argues in his contribution to "Data Worlds," researchers and teachers in nontechnical fields have for decades (and ever more blatantly since the 2008 financial crisis) lacked parity of any kind with the work of what are commonly described as STEM (science, technology, engineering, and mathematics) disciplines. On the topic of AI, in particular, "no public discussion," he writes, obeys a rule of "rough epistemic equality between technological and social knowledge." Paradoxically, this occurs even as so-called generative AI claims to reproduce skills and capacities central to the humanities, social sciences, and arts. But if the tech companies now vying for AI supremacy ignore the expertise of writers, artists, teachers, and humanists, the all-too-obvious truth is that they care little for public dialogue of any kind. Ex-Google CEO Eric Schmidt epitomized this mentality to a high degree when he claimed that no one in the US government was qualified to regulate AI—that only people in industry can get it "right" (see NBC News 2023).

As this special issue goes into production, the Writers Guild of America is on strike, partly to negotiate terms governing the use of automated writing. That is not because chatbots write good screenplays, but rather because studios can portray automated texts as first "drafts" so as to pay writers a lower rate for "revising" them, even though such "revision" is usually more laborious than writing from scratch.[20] This is but one of the many deceptions of "AI."

As the confabulations, simulacra, and mimicked biases of chatbots inundate the world's media ecosystem, as teachers and students are pushed to adopt "AI" in their classrooms or face imputations of Luddism or cluelessness, as the commercialization and training of ever-larger models squander energy and water at scale, as AI-adopting businesses pressure workers to produce more with less, and as gig workers in the global South endure industrial-speed toxicity in order to "mitigate" a supposedly automated technology, what becomes increasingly clear is that "AI" is not so much a technology as a political economy designed to concentrate power and profits in tech companies at the expense of everyone else. The novelist Ted Chiang (2023) makes a similar point in a recent *New Yorker* essay. AI isn't "dangerous" because it's likely to culminate in far-fetched doomer scenarios, he writes. Rather, AI is "dangerous inasmuch as it increases the power of capitalism."

Please join us in the shared project of forging an interdisciplinary community of critical practice that contributes to turning "AI" around.

**Notes**

1. Since *Critical AI*'s next issue (February 2024) is devoted to LLMs, here I offer the briefest footnote on this pervasive topic. For an important rejoinder to a much-discussed March letter (Future of Life Institute 2023) calling for a moratorium on "training AI systems more powerful than GPT-4," see Gebru et al. (2023). For an effort to mobilize understanding around the LLM hype cycle with humanist educators primarily in mind, see Goodlad and Baker (2023). As various contributors to "Data Worlds" discuss, data-driven ML and the related technique called "deep learning" (DL) were enabled by "big data," generated on the internet and through networked devices. "AI" today thus tends to involve data-driven ML at scale.
2. Golden's (2022) wiki defines *human in the loop* as the incorporation of "human intelligence in the process of creating and testing machine learning-based models."
3. The *Oxford English Dictionary* defines *critical thinking* as "the objective, systematic, and rational analysis and evaluation of factual evidence in order to form a judgment on a subject, issue, etc." Readers unfamiliar with *critique* or *critical theory* and curious about the specific meaning of *critical* in *Critical AI*, or critical AI studies, should be aware that the resonance is rooted in the practice of critical thinking and is never reducible to fault-finding or a thumbs-

down. *Critique* derives from the ancient Greek term for judgment and discernment; when taught in courses on *critical theory* it might include such thinkers as Aristotle, Immanuel Kant, G. W. F. Hegel, Karl Marx, Frantz Fanon, Hannah Arendt, or Michel Foucault. As Simon During (2020) has put it, critique has for centuries provided a "structuring condition" for the humanities, enabling them to mark out a cultural space distinct from business and partisan politics, even if "the humanities have by no means been consistently critical of dominant social values and institutions or, indeed, uninvolved in commerce and politics." The introduction to "Data Worlds" develops some ideas for a practice of critical AI studies; see also Raley and Rhee (2023) for an important special issue on the topic that was published too late to be discussed in this introduction.

4. We do not embrace Luddism in the conventional sense of being anti-technology; however, as Chiang (2023) points out, the historical Luddites were not actually resisting technology; rather, they resisted the economic injustice that use of technology was causing. In an article on AI doomerism, a dystopian discourse that forewarns of potential apocalypse due to out-of-control technologies, Weiss-Blatt (2023) notes how this hitherto esoteric outlook is being mainstreamed through, for example, opinion essays in the *New York Times* and *Time.* The doomer's far-fetched narrative of AI's ability to "wipe out humanity" reflects "a general preference for very amorphous, top-down . . . solutions." Paradoxically, doomers tend to be enthusiastic technologists, so that their speculations on nefarious, superintelligent technology represent another dimension of AI hype. Weiss-Blatt names Sam Altman (OpenAI's CEO) as a classic doomer, quoting an interview with ABC News in which he describes his anxieties over hypothetical AI attacks: "I try not to think about it too much. . . . But I have guns, gold, potassium iodide, antibiotics, batteries, water, gas masks from the Israeli Defense Force, and a big patch of land in Big Sur I can fly to." On Altman's tendency to blend this foreboding with optimistic boosterism—AI will, for example, invent fusion "and make the world wonderful"—see Harrison (2023). On Altman's use of such oxymoronic discourse to claim the need for regulation—but not the kind that regulates his own company's products—see, for example, Vincent (2023a, 2023b). For a recent scoop on OpenAI's intensive lobbying to water down the European Union's regulations, see Perrigo (2023b).
5. The phrase "more-than-human" signals openness to the flat ontologies that Bode and Goodlad discuss further in their introduction to this issue. While such terms deliberately complexify the ontology of the "human" in a technologically mediated and environmentally embedded world, it does not signal support for the eugenicist agenda of terms like *transhumanism* (see, for example, Torres 2023).
6. On AI's ongoing harms with respect to concentrated corporate power, see Kak and Wes (2023). On racial bias and inequality see, for instance, O'Neil (2016), Noble (2018), Buoalmwini and Gebru (2018), and Benjamin (2022). New work on the environmental harms of building and deploying AI systems and of so-called cloud computing is ongoing; for important examples, see Strubell, Ganesh, and McCallum (2019); Borning, Friedman, and Logler (2020); and Hogan (2021). For a new study on ChatGPT's water footprint, see Li et al. (2023). For a recent report on the harms of "generative AI," see Fergusson et al. (2023).
7. The word *robot* derives from the Czech term for a forced laborer and was introduced through Karel Čapek's 1920 science fiction play *R.U.R.* (in which manufactured human-like robots are synthesized from organic material). By contrast, *android*, to signify "an automaton resembling a human being," originated in the early modern era, with cited usages dating back to the seventeenth century. Literary and cultural touchstones for artificial and automated entities that predate the coinage of "AI" include the Greek legend of Pygmalion's creation of Galatea (narrated in Ovid's *Metamorphoses*); Shelley's *Frankenstein*; a wide discourse on automata (the construction of which dates back to the ancient world) that includes Schaffer's (1999: 127) discussion of how Enlightenment-era thinkers related "machinery viewed as human [to] humans managed as machines"; and Asimov's "positronic" robot stories, the earliest of which appeared in 1940.
8. On COMPAS, a system used for predicting criminal recidivism that was found to recommend longer sentences for black defendants than for white counterparts, see Angwin et al. (2016). See Silver et al. (2017) for DeepMind's original paper on AlphaZero, a system that used reinforcement learning trained itself to "superhuman" performance in chess with no access to domain knowledge beyond the rules of chess.
9. See Ziarek (2022) for a penetrating critique of AI from an Arendtian perspective.
10. The use of *AGI* as a synonym for cognate terms such as *strong* or *humanlike* intelligence predates the turn to deep learning by more than a decade (see, e.g., Goertzel and Pennachin 2007).
11. Many technologists use "artificial narrow intelligence" (ANI) to describe data-driven machine learning systems as "narrow" (limited in application) in contradistinction to the generality of human intelligence or "AGI." See Newfield in this issue for Smith's (2019) use of "reckoning" to describe the fine-grained computational analysis of large datasets at which machine learning excels. A different approach is to reject the label and speak of artificial "unintelligence," as does Broussard (2018). My own preference is instead to characterize "intelligence" pluralistically, and in more-than-human terms, while simultaneously avoiding the exaggerated anthropomorphisms that Broussard understandably seeks to curb. Machine intelligence varies considerably from human (and from all animal) intelligence. Moreover, *all* forms of intelligence are complexly embedded. Clearly, nonhuman life forms across the animal world, and arguably some plants, evince modes of intelligence that can enrich critical understanding (as Bode and I argue in our introduction to this issue). It follows that nuanced descriptions and analyses of machine intelligence (including data-driven "AI")—which avoid simplistic anthropomorphism, and which acknowledge the human-

generated data, labeling, design, reinforcement, and prompting that current technologies now require—represent an important avenue for critical AI studies.

12. For a compact survey of some of this midcentury terrain, see Lepore's (2020: 67–79) chapter on artificial intelligence, which notes how Wiener's *Cybernetics* (1948) compares biological nervous systems to machine systems.
13. McCarthy's (1997) definition ambiguously refers to an array of more-than-human entities and objects: "Varying kinds and degrees of intelligence occur in people, many animals and some machines." By contrast, Weizenbaum (1976), in rejecting the claim that human language understanding is conceptually reducible to "computer-manipulatable data structures," points to McCarthy to criticize the assumption that "life is what is computable and only that" (200). Weizenbaum (1976: 201) cites McCarthy as having said in 1973 (in defense of the slow progress of "AI" at that time), "The only reason we have not yet succeeded in simulating every aspect of the real world is that we have been lacking a sufficiently powerful logical calculus. I am currently working on that problem." See Silver et al. (2021: 1) for a paper from researchers at Google's DeepMind which hypothesizes "that intelligence, and its associated abilities" evolve in order to maximize reward such that "reward is enough to drive behaviour that exhibits abilities studied in natural and artificial intelligence, including knowledge, learning, perception, social intelligence, language, generalisation and imitation"—the implication being that the technique known as reinforcement learning, which has been successfully deployed in the automation of game-playing, is sufficient to reproduce the full range of (human) intelligence. As is common in some domains of AI research, the use of games as a proxy for real-world behaviors and spaces (and as an analogue for biological evolution) is never explored or questioned, even though games, unlike the real world, are predictable constructs that enforce particular rules and permissible moves in knowable spaces in order to cultivate the narrow set of tasks necessary for successful competition. Far from being a productive analogue for real-world complexity, game-playing falls short of that complexity.
14. As an example of how embodiment profoundly complexifies a brain-centric understanding of human and other animal behavior, consider recent research on what gastroenterologists (Carabotti et al. 2015: 203) call the "gut-brain axis," which posits "bidirectional communication" between the brain and "the central and the enteric nervous system, linking emotional and cognitive centers of the brain with peripheral intestinal functions" (including gut microbiota), which "is likely to have multiple effects on affect, motivation, and higher cognitive functions."
15. Many scholars are researching the origins of "AI" origins in statistical innovations and biometric and eugenicist pseudoscience, e.g., Mhlambi (2020), Chun (2021), Crawford (2021), Gilbert (2023), and Goodlad (2020, 2022). As Gould puts it (1996), statistically normed metrics such as IQ tests promote the "abstraction of intelligence" as a "single, quantifiable entity located in the brain" and then use these metrics to assert the biological inferiority of marginalized races, classes, or sexes (21).
16. As Birhane argues (2020: 395–96), the brute force colonialism of the past is now succeeded by algorithmic variants that inject the "values, norms, and interests of Western societies" into Africa, where the lack of attention to social contexts has been especially damaging to the health sector.
17. For an example of such a paper, see Bubeck et al. (2023).
18. For the most thorough account of the harms of so-called generative AI thus far, see Fergusson et al. (2023). Raji et al.'s (2022: 962) "failure taxonomy" begins with tasks that are conceptually or practically impossible; proceeds to engineering failures (of design, implementation, and/or missing safety features); discusses postdeployment failures (lack of robustness, vulnerability to adversarial attacks, and/or unanticipated interactions); and concludes with communication failures involving the falsification, overstatement, and/or misrepresentation of capabilities. To be clear, a systematic address of such deep-seated failures could significantly help to curb long-term risks. The point is not to pit current harms against long-term safety, but rather to prevent narrowly defined and far-fetched risks from diverting attention from actually existing harms and broad regulatory goals.
19. See also Gebru et al. (2023), Lavigne in this special issue, and the Center for Artistic Inquiry and Reporting's (2023) open letter, which argues that the datasets that make art generators possible contain "millions upon millions of copyrighted images, harvested without their creator's knowledge, let alone compensation or consent"—in effect, "the greatest art heist in history."
20. For commentary on the strike, see Merchant (2023) and Harris (2023).

**Works Cited**

Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. "Persistent Anti-Muslim Bias in Large Language Models." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306.

Altman, Sam. 2023. "Planning for AGI and Beyond." OpenAI, February 24. https://openai.com/blog/planning-for-agi-and-beyond.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*, May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arendt, Hannah. 1951. *The Origins of Totalitarianism*. New York: Harcourt.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. New York: Association for Computing Machinery.

Bender, Emily M., and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. https://aclanthology.org/2020.acl-main.463.

Benjamin, Ruha. 2022. *Viral Justice: How We Grow the World We Want*. Princeton, NJ: Princeton University Press.

Birhane, Abeba. 2020. "Algorithmic Colonization of Africa." *Scripted: A Journal of Law, Technology, and Society* 170, no. 2 (August): 389–409.

Borning, Alan, Batya Friedman, and Nick Logler. 2020. "The 'Invisible' Materiality of Information Technology." *Communications of the ACM* 63, no. 6 (May): 57–64.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: MIT Press.

Bubeck, Sébastien, et al. 2023. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." *arXiv*, April 13. https://arxiv.org/abs/2303.12712.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 83, no. 1: 1–15.

Carabotti, Marilia, Annunziata Scirocco, Maria Antonietta Maselli, and Carola Severi. 2015. "The Gut-Brain Axis: Interactions between Enteric Microbiota, Central and Enteric Nervous Systems." *Annals of Gastroenterology* 28, no. 2 (April–June): 203–9.

Center for Artistic Inquiry and Reporting. 2023. "Restrict AI Illustration from Publishing: An Open Letter." CAIR, May 2. https://artisticinquiry.org/AI-Open-Letter.

Chiang, Ted. 2017. "Silicon Valley Is Turning into Its Own Worst Fear." *Buzzfeed News*, December 18. https://www.buzzfeednews.com/article/tedchiang/the-real-danger-to-civilization-isnt-ai-its-runaway.

Chiang, Ted. 2023. "Will A.I. Become the New McKinsey?" *New Yorker*, May 4.

Chun, Wendy Hui Kyong. 2021. *Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition*. Cambridge, MA: MIT Press.

Costanza-Chock, Sasha, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–83. Seoul: ACM.

Crawford, Kate. 2021. *Atlas of AI*. New Haven, CT: Yale University Press.

Davis, Martin. 2000. *The Universal Computer: The Road from Leibniz to Turing*. Boca Raton, FL: Norton.

During, Simon. 2020. "Are the Humanities Modern? After Latour." In *Latour and the Humanities*, edited by Rita Felski and Stephen Muecke, 225–48. Baltimore: Johns Hopkins University Press.

Fergusson, Grant, Calli Schroeder, Ben Winters, and Enid Zhou, eds. *Generating Harms: Generative AI's Impact and Paths Forward*. Electronic Privacy Information Center, May. https://epic.org/wp-content/uploads/2023/05/EPIC-Generative-AI-White-Paper-May2023.pdf.

Future of Life Institute. 2023. "Pause Giant AI Experiments: An Open Letter." March 22. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gebru, Timnit, and Émile P. Torres. 2023. "SaTML 2023—Timnit Gebru—Eugenics and the Promise of Utopia through AGI." YouTube. Posted by Nicholas Papernot, February 16. https://www.youtube.com/watch?v=P7XT4TWLzJw.

Gebru, Timnit, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell. 2023. "Statement from the Listed Authors of Stochastic Parrots on the 'AI Pause' Letter." Distributed AI Research Institute, March 31. https://www.dair-institute.org/blog/letter-statement-March2023/.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models." In *Findings of the Association for Computational Linguistics: EMNLP*, 3356–69. https://aclanthology.org/2020.findings-emnlp.301/.

Gilbert, Pamela. 2023. "Common Sense, AI, and the Whiteness of Affect." Paper presented at "Victorian 'Artificial Intelligence': A Call to Arms," New Brunswick, NJ, April 6.

Goertzel, Ben, and Cassio Pennachin. 2007. *Artificial General Intelligence*. Berlin: Springer.

Golden. 2022. "Human-in-the Loop." Edited May 22. https://golden.com/wiki/Human-in-the-loop-Y8R3PM/activity.

Goodlad, Lauren M. E. 2020. "A Study in Distant Reading: Genre and the Longue Dureé in the Age of AI." *Modern Language Quarterly* 81, no. 4 (December): 491–525.

Goodlad, Lauren M. E. 2022. "Victorian 'Artificial Intelligence': or How George Eliot's Fiction Helps Us to Understand Statistical Modelling." Sally Ledger Memorial Lecture, London, April 7.

Goodlad, Lauren M. E., and Samuel Baker. 2023. "Now the Humanities Can Disrupt AI." *Public Books*, February 20. https://www.publicbooks.org/now-the-humanities-can- disrupt-ai/.

Gould, Stephen Jay. 1996. *The Mismeasure of Man*. 2nd ed. New York: Norton.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Harris, Mary. 2023. "Who's Afraid of A.I.?" Interview with Meredith Whittaker. *What's Next: TBD*. Podcast audio, May 12. https://slate.com/podcasts/what-next-tbd/2023/05/why-artificial-intelligence-needs-regulation-now.

Harrison, Maggie. 2023. "Sam Altman Says AGI Will Invent Fusion and Make the World Wonderful." *Futurism*, May 9. https://futurism.com/sam-altman-agi-fusion-world.

Hofstadter, Douglas R. 1996. *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought.* New York: Basic Books.

Hogan, Mél. 2021. "The Data Center Industrial Complex." In *Saturation: An Elemental Politics*, edited by Melody Jue and Rafico Ruiz, 283–305. Durham, NC: Duke University Press.

Hundt, Andrew, William Agnew, Vicky Zeng, Severin Kacianka, Matthew Gombolay. 2022. "Robots Enact Malignant Stereotypes." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 743–56. Seoul: ACM.

Ingram, David. 2023. "ChapGPT Is Powered by These Contractors Making $15 an Hour." *NBC News*, May 6. https://www.nbcnews.com/tech/innovation/openai-chatgpt-ai-jobs-contractors-talk-shadow-workforce-powers-rcna81892.

Irani, Lilly C., and M. Six Silberman. 2013. "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk." In *Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems*, 611–20. Paris: SIGCHI.

Kak, Amba, and Sarah Myers Wes. 2023. "AI Now 2023 Landscape: Confronting Tech Power." AI Now Institute, April 11. https://ainowinstitute.org/2023-landscape.

Lanier, Jaron. 2023. "There Is No A.I." *New Yorker*, April 20.

Lepore, Jill. 2020. *If Then: How the Simulmatics Corporation Invented the Future*. New York: Norton.

Lighthill, James. 1973. "Artificial Intelligence: A General Survey." Chilton Computing and UKRI Science and Technology Facilities Council. http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." Paper presented at ACL 2022, Dublin, May.

Li, Pengfei, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2023. "Making AI Less 'Thirsty': Uncovering the Secret Water Footprint of AI Models." *arXiv*, April 6. https://arxiv.org/abs/2304.03271.

Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York: Pantheon Books.

McCarthy, John. 1997. "What Is Artificial Intelligence?" Lecture presented at Stanford University, Stanford, CA.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and C. E. Shannon. 2006. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955." *AI Magazine* 27, no. 4 (winter): 12–14.

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115–33.

Merchant, Brian. 2023. "Your Boss Wants AI to Replace You: The Writer's Strike Shows How to Fight Back." *Los Angeles Times*, May 11.

Mhlambi, Sabelo. 2020. "From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance." Carr Center Discussion Paper Series. Cambridge, MA: Harvard University.

NBC News. 2023. "Fmr. Google CEO Says No One in Government Can Get AI Regulation 'Right.'" Posted by NBC News. YouTube, May 15. https://www.youtube.com/watch?v=oewQC4wDTYQ.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Perrigo, Billy. 2023a. "OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic." *Time*, January 18. https://time.com/6247678/openai-chatgpt-kenya-workers/.

Perrigo, Billy. 2023b. "Exclusive: OpenAI Lobbied the EU to Water Down AI Regulation." *Time*, June 23. https://time.com/6288245/openai-eu-lobbying-ai-act/.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. "AI and the Everything in the Whole Wide World Benchmark." Paper presented at the 35th Conference on Neural Information Processing Systems, online, August 20.

Raji, Inioluwa Deborah, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. "The Fallacy of AI Functionality." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–72. Seoul: ACM.

Raley, Rita, and Jennifer Rhee, eds. 2023. "Critical AI: A Field in Formation." Special issue, *American Literature* 95, no. 2.

Schaffer, Simon. 1999. "Enlightened Automata." In *The Sciences in Enlightened Europe*, edited by William Clark, Jan Golinski, and Simon Schaffer, 126–68. Chicago: University of Chicago Press.

Sejnowski, Terrence J. 2018. *The Deep Learning Revolution*. Cambridge, MA: MIT Press.

Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton. 2021. "Reward Is Enough." *Artificial Intelligence* 299: 103535. https://doi.org/10.1016/j.artint.2021.103535.

Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." In *Proceedings of the 57th Annual Meeting of the Association of Computational Linguistics*, 3645–50. Florence: ACL.

Turing, Alan M. 1950. "Computing Machinery and Intelligence. *Mind* 49, no. 236: 433–60.

Torres, Émile P. 2023. "Longtermism and Eugenics: A Primer." *Truthdig*, February 4. https://www.truthdig.com/articles/longtermism-and-eugenics-a-primer/.

Vincent, James. 2023a. "The Senate's Hearing on AI Regulation Was Dangerously Friendly." *The Verge*, May 19. https://www.theverge.com/2023/5/19/23728174/ai-regulation-senate-hearings-regulatory-capture-laws.

Vincent, James. 2023b. "OpenAI Says It Could 'Cease Operating' in the EU if It Can't Comply with Future Regulation." *The Verge*, May 25. https://www.theverge.com/2023/5/25/23737116/openai-ai-regulation-eu-ai-act-cease-operating.

von Neumann, John. 1958. *The Computer and the Brain*. New Haven, CT: Yale University Press.

Weidinger, Laura, et al. 2022. "Taxonomy of Risks Posed by Language Models." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–29. Seoul: ACM.

Weiss-Blatt, Nirit. 2023. "The AI Doomers' Playbook." *Techdirt*, April 14. https://www.techdirt.com/2023/04/14/the-ai-doomers-playbook/.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. New York: Freeman.

Whittaker, Meredith. 2021. "The Steep Cost of Capture." *Interactions* 28, no. 6: 50–55.

Wiener, Norbert. 1948. *Cybernetics or Control and Communication in the Animal and the Machine.* Cambridge, MA: MIT Press.

Ziarek, Ewa Pøonowska. 2022. "Against Digital Worldlessness: Arendt, Narrative, and the Onto-politics of Big Data/AI Technologies." *Postmodern Culture* 32, no. 2 (January). https://muse.jhu.edu/pub/1/article/864899.