Beyond Chatbot-K: On Large Language Models, "Generative AI," and Rise of Chatbots: An Introduction

Lauren M. E. Goodlad and Matthew Stone

Abstract This essay introduces the history of the "generative AI" paradigm, including its underlying political economy, key technical developments, and sociocultural and environmental effects. In concert with this framing it discusses the articles, thinkpieces, and reviews that make up part 1 of this two-part special issue (along with some of the content for part 2). Although Large Language Models (LLMs) are marketed as scientific wonders, they were not designed to function as either reliable interactive systems or robust tools for supporting human communication or information access. Their development and deployment as commercial tools in a climate of reductive data positivism and underregulated corporate power overturned a long history in which researchers regarded chatbots as "misaligned" affordances for safe or reliable public use. While the technical underpinnings of these much-hyped systems are guarded as proprietary secrets that cannot be shared with researchers, regulators, or the public at large, there is ample evidence to show that their development depends on the expropriation and privatization of human-generated content (much of it under copyright); the expenditure of enormous computing resources (including energy, water, and scarce materials); and the hidden exploitation of armies of human workers whose low-paid and high-stress labor makes "AI" seem more like human "intelligence" or communication. At the same time, the marketing of chatbots propagates a deceptive ideology of "frictionless *knowing*" that conflates a person's ability to leverage a tool for producing an output with that person's active understanding and awareness of the relevant information or truth claims therein. By contrast, the best digital infrastructures for human writing enable human users by amplifying and concretizing their interactive role in crafting trains of contemplation and rendering this situated experience in shareable form. The essay concludes with reflections on alternative pathways for developing AI—including communicative tools—in the public interest.

Keywords: large language models, generative AI, chatbots, data positivism, artificial general intelligence (AGI), design justice principles

About midway through an opinion essay in the *New York Times*, Yuval Harari, Tristan Harris, and Aza Raskin (2023)—authors who share an interest in the future of technology—wax magniloquent as they build their discussion of "AI" to a weighty crescendo: "The specter of A.I. has haunted humanity since the mid-20th century, yet until recently it has remained a distant prospect, something that belongs in sci-fi more than serious scientific and political debates. It is difficult for human minds to grasp the new capabilities of GPT-4 and similar tools, and it is even harder to grasp the exponential speed at which these tools are developing more advanced and powerful

capabilities." The "AI" technologies in question are large language models (LLMs) like OpenAI's GPT-4 (now incorporated into Microsoft's Bing), Google's Gemini, and Anthropic AI's Claude 3—systems that also go by the names *generative AI*, *foundation models*, and *frontier models*. We will return to this terminological array but for now pause to emphasize how Harari, Harris, and Razkin imbue these advances in the modeling of data with existential import and scientific mystique ("difficult for human minds to grasp"). The result, as mathematician Noah Giansiracusa (2023) puts it, is "reckless exaggeration." For example, while LLMs have grown exponentially larger in terms of the size of the data sets and the number of parameters, the performance boost earned from each round of costly and resource-intensive enlargement has in fact substantially decreased. These "diminishing returns" (Thompson et al. 2021) have led many experts, including OpenAI's CEO, to opine that "further progress will not come from making models bigger" (Knight 2023b). At the same time, some of the most significant improvements since the 2020 release of GPT-3 have involved industrial-scale human feedback. Indeed, according to a recent Google Research paper (Dzieza 2023), "millions" of human workers are now hired to improve these ostensibly automated systems "with the potential to become 'billions.'" Thus, the speed of "AI" development varies widely across tasks—but in no case can the pace be properly described as "exponential." Finally, the proposition that "human minds" cannot grasp the "new capabilities of GPT-4" marks a perplexing depreciation of human intelligence that threatens to undermine necessary discussions about technology before they have even begun.

As they continue spinning a yarn about the "specter" of AI, Harari and colleagues describe language ("the word") in a quasi-theistic idiom,[1] as if the advent of "AI" were a genesis narrative inflected by prophecy from the Book of Revelation. "In the beginning was the word. Language is the operating system of human culture. From language emerges myth and law, gods and money, art and science, friendships and nations and computer code." Notably, the metaphor of language as "operating system" presupposes an understanding of "human culture" as an information machine working to the dictates of a central authority. As the sociolinguist Britta Schneider writes in her thinkpiece in this special issue, "Western cultures often regard language as an immaterial phenomenon," neglecting its emergence from "bodily and material coordinating practices" that are jointly enacted "through vocal cords, hands, faces, or written signs." Harari and colleagues thus set aside the diverse and historically situated origins and trajectories of human signifying capacities to conjure a single telos that (despite their contrary claim) seems ripped from the pages of science fiction:

> A.I.'s new mastery of language means it can now hack and manipulate the operating system of civilization. By gaining mastery of language, A.I. is seizing the master key to civilization, from bank vaults to holy sepulchers.

[1] See Keane and Shapiro 2023 for an illuminating discussion of what the authors call "AI godbots"—a term devised to express the ability of chatbots to exploit a human tendency to "impute divinity to inexplicable processes," while simultaneously trying "to get the gods to talk to us."

> . . . Simply by gaining mastery of language, A.I. would have all it needs to contain us in a *Matrix*-like world of illusions. . . . A curtain of illusions could descend over the whole of humanity. . . .
>
> Democracy is a conversation, conversation relies on language, and when language itself is hacked, the conversation breaks down, and democracy becomes untenable. If we wait for chaos to ensue, it will be too late to remedy.

As this introduction will clarify, while today's language models are impressive and selectively useful, they have not "mastered language" in any fundamental sense.[2] Below we will elaborate how LLMs (and their multimodal successors) are data-driven statistical models that initially emerged from research techniques in speech-to-text transcription and machine translation. Though the idea for such artificial "neural networks" first appeared eighty years ago (McCulloch and Pitts 1943), the technique's potential dramatically improved in the 2010s due largely to advances in hardware and the historically unprecedented stores of human-generated language data concentrated on the internet during the preceding decade. Yet while the affordances of state-of-the art LLMs may "surprise" us (as computer scientist Sam Bowman puts it in his contribution to part 2 of this special issue), these systems are in no position to "hack" civilization—though, as we will see, generative AI poses many harms including the potential for bad actors to hijack these new technologies for malicious use or their own advantage.

Recall that in *The Matrix* (1999), a fictional artificial intelligence creates a world-within-a-world so that human beings unknowingly inhabit a metaverse-like simulation while they function as energy resources for their machine captors. By contrast, the idea of *language* as the vehicle for a "*Matrix*-like world of illusions" ostensibly nods to Marxist notions of an alienating false consciousness or to poststructuralist propositions about how human subjectivities take shape through the habituating practices of modern institutions. Unlike Harari and colleagues' far-fetched warnings of machine takeover, these theoretical meditations on the interaction between cultural narratives and material structures typically focus on political economies, discursive regimes, and relations of power. In the case of "AI," the point could be to examine how, and for what purposes, particular automated systems have been designed, optimized, and deployed—processes that, in the case of generative AI, have tended to heavily rely on the labor of poorly compensated human workers. As the computer scientist Rediet Abebe and the economist Maximilian Kasy (2021) put it, "Whose goals count?"[3]

In *Confronting Tech Power* (Kak, West, and Whittaker 2023), researchers from the AI Now Institute illustrate that approach. The idea of LLMs as foundation models, they note, originated at Stanford University in concert with the launch of the Center for

---

[2] On this point see also Bender 2022, a response to a different *New York Times* feature (S. Johnson and Iziev 2022) that argued misleadingly that OpenAI's GPT-3 (a precursor to ChatGPT) had "mastered language." We would also challenge the patriarchal and authoritarian imagery inherent in the very idea of "mastering" language.

[3] For a relevant discussion of the reactionary cooptation of the film's metaphor of taking the "red pill" for an array of rightwing conspiracy theories, see Chun 2021: 29–34.

Research on Foundation Models. By portraying proprietary commercial systems like ChatGPT as "foundation models" for future research, Stanford's coinage plays to the strengths of the world's largest tech companies and their elite academic partners—institutions that already possess formidable advantages. The consequent determination to dominate research agendas, monopolize markets, release untested tools, destabilize long-standing norms, and capture regulatory regimes does indeed point to antidemocratic tendencies and effects. But rather than play up the wizardry of "AI," *Confronting Tech Power* documents a familiar problem of too much control in the hands of too few—comparable to the monopolization of railroads and oil during the last Gilded Age.

As if to confirm this perspective, Judy Estrin (2023), the former chief technology officer of Cisco, offers a sobering commentary quite unlike Harari and colleagues' fears of "godlike" AI. Arguing that "AI" in effect stands for "*authoritarian* intelligence," Estrin writes in *Time*:

> I have never had such mixed feelings about technological innovation. In stark contrast to the early days of internet development, when many stakeholders had a say, discussions about AI and our future are being shaped by leaders who seem to be striving for absolute ideological power. . . . The hubris and determination of tech leaders to control society is threatening our individual, societal, and business autonomy.
>
> What is happening is not just a battle for market control. A small number of tech titans are busy designing our collective future, presenting *their* societal vision, and specific beliefs about *our* humanity, as the only possible path. Hiding behind an illusion of natural market forces, they are harnessing their wealth and influence to shape not just productization and implementation of AI technology, but also the research.

As examples of what Estrin may have in mind, consider the appeals to the public that "tech titans" have made since OpenAI's release of ChatGPT in November 2022. For instance, "Pause Giant AI Experiments" is a March 2023 letter disseminated by the Future of Life Institute (2023)—an organization of which Elon Musk, a prominent signatory of the letter, both board member and major sponsor.[4] Although the long list of signatures included many people who simply seized an opportunity to support regulation, Future of Life's letter foreground the priorities of tech leaders who, like Musk himself, dramatize the supposedly existential risks of the very products they are vying to capitalize and commercialize. Indeed, so far from heeding his own call for a "pause" (Thorbecke 2023), Musk has continued to boast about his plans to rival OpenAI with a new company that will plumb "the true nature of the universe."[5]

[4] On the abrupt April 2024 closing of Oxford University's Futures of Humanity Institute, an organization with a similar mission, in the wake of controversies over the alleged racism of founder and philosopher Nick Bostrom, see, for example, Anthony (2024).

[5] Many commentators use the informal term "doomer" to describe the dystopian rhetoric of elite technologists. As Nirit Weiss-Blatt (2023) argues, this once esoteric Silicon Valley narrative of technology's ability to "wipe out humanity" is another route to upholding "top-down" solutions.

An even more central player in generative AI is Microsoft: the software giant's 49 percent stake in OpenAI—the latest investment in a partnership that began in 2016 (see, e.g., Statt 2016)—marks both a bid to challenge Google's lucrative domination of web search and a strategy to augment Microsoft's sale of web services (a profitable business in which Microsoft has been gaining ground on the industry leader, Amazon). Microsoft CEO Satya Nadella described his determination to harness OpenAI's GPT-4 to his company's Bing search engine as a desire "to make Google dance" (quoted in Steven Levy 2023). But Nadella's C-suite bravado did not stop Microsoft founder Bill Gates from joining the CEOs of Open AI, Anthropic AI (in which both Amazon and Google hold major stakes), and DeepMind (a Google subsidiary) in signing the Center for AI Safety's May 2023 "Statement of AI Risk."[6] The May statement argues that an AI-driven "risk of extinction" is a "global priority" comparable to that of nuclear war, yet remarkably makes no mention of climate change. Once again, the world's most powerful tech leaders collaborated in sensationalistic messaging about "AI" while positioning themselves as the bearers of paramount priorities and unchallengeable authority.

In a further installment of this ongoing storyline, AI "godfather" Geoffrey Hinton, another prominent signatory of the May 2023 statement, told the *New York Times* that he had resigned from Google so as freely to "criticize" the tech industry's premature release of new products (Metz 2023). But within days of expressing some reasonable concerns about an underregulated industry,[7] Hinton (2023) clarified that his actual goal was not to rein in the "race to deploy" but simply to join the chorus of leaders already prophesying "existential" dangers.[8] Queried by CNN's Jake Tapper as to what should be done about these supposed threats to the human species, the scientist replied that he wasn't sure if any "solution" is possible. Asked if he regretted his failure to support Google researchers like Timnit Gebru—one of two AI ethicists who lost their jobs at Google because of a research paper that warned of LLMs' dangers (Bender et al. 2021; Hao 2020)—Hinton replied that these colleagues' concerns were less "existentially

---

Paradoxically, doomerism typically works hand in hand with enthusiastic techno-optimism so that it represents a strain of elite technologist discourse, sometimes referred to as "doomer-boosterism." In an early discussion of the phenomenon, Ted Chiang (2017) argues elegantly that Silicon Valley's fears of rogue AI mirror the industry's own antisocial features. But, for an example of unadulterated AI boosterism and techno-utopianism, see Elizabeth Spiers's (2023) opinion essay on investor Marc Andreesen's self-styled techno-optimist manifesto, "A Tech Overlord's Horrifying, Silly Vision for Who Should Rule the World"; for a biting take on the phenomenon of technocratic hubris more generally, see Rushkoff 2023. As of March 2024, Musk is suing OpenAI (a company he had helped to fund, and one that retains its technically nonprofit structure) for abandoning its supposedly beneficial research mission; OpenAI is countering that Musk himself sought to control the company absolutely (see, e.g., Heath and Lopatto 2024).

[6] "Statement on AI Risk," Center for AI Safety, May 2023, https://www.safe.ai/statement-on-ai-risk.

[7] According to the *Times*'s initial report (Metz 2023), Hinton was concerned about a potential flood of fake media, disruptions to the job market, and the possibility that as Google "races to deploy" a product comparable to Microsoft's OpenAI-powered products, the two "tech giants" could lock themselves "in a competition that might be impossible to stop."

[8] As Google researcher François Chollet (2023) points out, "panic" over "imminent AGI" has recurred since 2013 following DeepMind's programs for playing Atari and later Go. See also Goldman 2023 on debates among AI's so-called godfathers over "doomer" rhetoric.

serious" than his own. His interview thus confirmed a pattern in which fears of a far-fetched AI catastrophe displace recognition of actually existing harms to people and the planet. Asked if tech companies themselves can be trusted to self-regulate despite their focus on profits, Hinton demurred; the same companies releasing these models were also "the most likely" to keep AI "under control," he asserted.

Behind such muddled thinking is a vindication of Estrin's claim that grandiose tech leaders are advancing their own "societal vision" as if it were the "only possible path." As high-profile technologists urge regulators to focus on extraordinary scenarios, they simultaneously insist that the industry's aggressive commercialization of untested "AI" is a path to progress that must be entrusted to the industry's own oversight. This incoherent messaging can be surreptitious (as in CEO Altman's behind-the-scenes lobbying of the European Union to insulate OpenAI's programs from the same regulatory interventions that he was publicly calling for [Perrigo 2023a]), downright confounding (as in Hinton's on-again/off again "criticism" of the industry), or baldly declared—as in the claim of Eric Schmidt, former CEO of Google, that "no one in government" has the necessary expertise to regulate AI (NBC News 2023).

In their reply to the March 2023 letter, Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell (2023), a group of AI ethicists whose research the letter had cited, argued that any serious effort to regulate so-called AI must acknowledge real and present harms and focus on "transparency, accountability and preventing exploitative labor practices." To quote coauthor Mitchell, "Ignoring active harms . . . is a privilege" that most people "don't have" (Coulter 2023)—including the inability to ignore (in the words of Alex Hanna and Emily M. Bender [2023]) "wrongful arrests, an expanding surveillance dragnet, defamation and deepfake pornography." In fact, researchers in AI ethics, including many pioneering scholars of color, have been sounding the alarm about data-driven machine learning (ML) for quite some time.[9]

In her thinkpiece in this special issue, Annette Vee, a scholar of digital composition and literacy, adopts the terms *technical debt* and *moral hazard* to describe how dominant tech companies shape the industry's political economy when they shift the material costs of short-term decision-making onto workers, creative people, minoritized cultures, underresourced regulators, and the public at large. These underlying tensions between big tech imperatives and public interest—as the

[9] As Lorena O'Neil (2023) notes, female scholars of color have been especially prominent in warning of potentially "disastrous" effects in "magnifying biases," "stripping out" necessary contexts, and making decisions about people who lack "the choice to opt out." For important pioneering work for a critical AI studies perspective, see Cathy O'Neil's (2016) warning of the antidemocratic and inequitable effects of algorithmic decision-making; Safiya Umoja Noble's (2018) analysis of how search engines prioritize the popularity of websites over their legitimacy; Joy Buolamwini and Timnit Gebru's (2018) research on the intersectional discriminatory effects to which facial recognition systems subject women of color; Ruha Benjamin's (2019) case for a "Jim Code" that embeds racial biases into diverse digital systems; Emma Strubell, Ananya Ganesh, and Andrew McCallum's (2019) documentation of the energy costs of training LLMs; legal scholar Frank Pasquale's (2020) emphasis on how automated systems are marketed as cost-effective alternatives to professional expertise; and Emily M. Bender et al.'s (2021) research on the "dangers" of models trained on proprietary data sets that are too large (and too secretive) to document and hold to account.

philosophers Jan-Christoph Heilinger and Hendrik Kempt note in their thinkpiece—are sometimes framed as if they marked a regrettable impasse between promoters of "safety" who focus on "existential" threats and promoters of "ethics" who focus on actually existing harms. The framing of such an impasse favors the "safety" camp, who may portray critics as out-of-touch skeptics who do not recognize the technology's extraordinary powers.[10] In reality, of course, "ethical" demands for corporate transparency, mandatory audits, robust regulatory standards, liability for harms, and the enactment of laws that protect workers and creative people do not undermine the public's ability to address more speculative dangers. By way of intervention, Heilinger and Kempt schematize AI ethics across conceptual, substantive, and procedural dimensions. While the industry favors technological solutions to technological risks, they conclude, an AI ethics that matters must push for public oversight that includes meaningful input from affected people and communities.

In an object lesson of corporate priorities at work, OpenAI took center stage almost a year to a day after ChatGPT's launch, when the company's little-discussed board summarily fired its high-profile CEO, Sam Altman, only to reverse itself a few days later. One notable feature of this melodrama was the tension between Altman and board member and chief scientist Ilya Sutskever, who, according to the *Atlantic*, had devised his own ritual burning of an effigy of "malign" AI (Hao and Warzel 2023). As against Sutskever's doomer-inflected "safety" orientation, Altman himself (a Stanford dropout who often talks down the value of education) is, according to the *New York Times*, a leader who, with "little engineering training," is "driven by a hunger for power" (Mickle et al. 2023). During secretive and tense negotiations, the two female board members who had supported the ouster were replaced by several men, including Lawrence Summers, an economist and former treasury secretary, infamous for impugning women's biological capacity to excel at math and science (Goldenberg 2005; Hauser 2023). As *Los Angeles Times* tech columnist Brian Merchant (2023b) observed, it had taken only forty-eight hours for the priorities of investors like Microsoft to "obliterate" the board's governance and its ability to address alleged "safety" concerns.[11]

---

[10] For an excellent review essay that notes that AI's boosters have a vested interest both in projecting the the "future-focused gloss of science fiction" onto their enterprise and in "imply[ing] subtly to the comfortable that the disruptive social impacts of AI systems are safely in the future. A focus on AI safety satisfies these ideological goals admirably," see Stark 2023: 368.

[11] See Mandaro, Mascarenhas, and Palazzolo 2024 on Altman's March 2024 reinstatement to the board. In their in-depth coverage of the November ouster, Karen Hao and Charlie Warzel (2023) report that the ostensibly startling action was the culmination of a long-brewing "power struggle" between "two ideological extremes": one born from "Silicon Valley techno-optimism, energized by rapid commercialization; the other steeped in fears that AI represents an existential risk to humanity." The future of this consequential technology, they wrote, "is being determined by an ideological fight between wealthy techno-optimists, zealous doomers, and multibillion-dollar companies." See also the *Washington Post*'s interview with the director of the Revolving Door watchdog group: "There is no greater indication that OpenAI is unserious about the interests of humanity," said the director, than their elevation of Larry Summers," whose ascent to its board "should accelerate concerns that AI will be bad for all but the richest and most opportunistic amongst us" (Verma, Tiku, and De Vynck 2023). On Sutskever's subsequent departure from OpenAI in May 2024, see, for example, R. Metz (2024).

To be sure, some of the most harmful modes of "AI" are predictive systems for tasks such as facial recognition, detection of potential benefits fraud, or algorithmic pricing, tasks that do not directly depend on the new "generative" technologies (see, e.g., Raji et al. 2020; Heikkila 2022; West 2023). In focusing this special issue on the role LLMs play in capturing the zeitgeist, we are motivated partly to explore how the November 2022 launch of ChatGPT spurred a historic tide of popular interest in, and concern about, something called "generative AI," which was now taking the primary form of chatbots. This hype cycle has been sustained ever since through much-ballyhooed follow-ups including Microsoft's leveraging OpenAI's chat technology in February 2023 and since; Alphabet's haste to respond by harnessing their Bard chatbot to Google search in the very same month (Pichai 2023); the steady release of competitor LLMs, including Alphabet's Gemini (which, within days of its December 2023 release, was found to have "hyped up" its promotional video through misleading edits [Edwards 2023a]), Alphabet's controversial introduction of "AI Overview" into Google Search in May 2024 (Grant 2024), and Anthropic's Claude 3 Opus (released in March 2024); OpenAI's periodic updates to its chat products (including the controversial introduction and subsequent pause of a voice feature widely perceived to mimic Scarlett Johansson's performance of a fictional digital assistant in Spike Jonze's *Her* [2013] [e.g., Tiku 2024][see also Stone, Goodlad, and Sammons in this issue]); coverage of sundry downstream products for generating text, images, computer code, humanlike vocalization, music, and video (including widespread concerns over "deepfakes"); major labor actions turning partly on AI (such as the historic Writers Guild of America and SAG-AFTRA strike in 2023); and industry-led publications (e.g., Bubeck et al. 2023) documenting the supposed emergence of human- or superhuman-level artificial general intelligence (AGI).

Media commentary on generative AI criss-crosses domains, venues, and technical or political standpoints. One strain of prominent media fare showcases tech leaders' diverse fusions of the booster-doomer paradigm. For example, DeepMind cofounder Mustafa Suleyman, promoting a new book and start-up centered on "empathetic AI" (Inflection 2024), purveys a more auspicious brand of speculation than the standard doomer narrative. On the one hand, he told *AP News* that AI will deliver an "era of radical abundance" by endowing "everyone on the planet" with "broadly equal access to intelligence" (Liedtke 2023). On the other hand, Suleyman's (2023) book warns that AI could make us vulnerable to algorithmic "systems beyond our control." Fortunately, he told *Wired*, the necessary regulation to preclude such harms "is just going to be another component" to a system of internet governance that, as Suleyman regards it, is already "pretty good." Microsoft founder Gates, meanwhile—a perennial fixture of AI hype cycles—followed his affirmation of the grave risks of human extinction in May 2023, with a sunnier outlook in November. According to the UK *Independent*, Gates speculates that "AI" could usher in a three-day work week, enabling humans to "do a lot less work to get by" (Fletcher 2023). Unmentioned by Gates or the *Independent* is that bullish techno-optimists have predicted leisure and plenty for the masses for centuries—even as the very same soothsayers embrace free market doctrines that leave workers too precarious to claim a fair share of their productivity gains (see, e.g., Giridharadas 2018; Stoller 2019; Merchant 2023a; Rushkoff 2022).

Of course, not all mainstream media toe the industry line. Writing in the *New York Times*, tech columnist Julia Angwin (2023) reports that proprietary generative systems have begun to damage the internet's cultural content (see also Hsu and Thompson 2023). *New York*'s John Herrman (2024) writes incisively about the dilemmas of "unscoped" (general-purpose) chatbots like ChatGPT and Gemini, which are doomed repeatedly to stumble over culture wars. *Critical AI*'s contributors to this special issue draw selectively on strong reporting in a range of venues including *Time*, the *Washington Post*, the *Markup*, the *Verge*, *Logic(s)*, and the recently launched *404 Media* website. By contrast, many prominent business publications, including industry-friendly academic venues, confirm the narrative of a major technological revolution, with little or no reference to the impact of hype cycles or speculative bubbles. According to Stanford's Center for Human-Centered Artificial Intelligence, AI will "transform teaching and learning" (Chen 2023). The influential consultancy firm McKinsey & Company predicts that generative AI will tackle the "biggest burdens" in health care (Bhasker et al. 2023), while the *Harvard Business Review* anticipates both that generative AI will "disrupt" the "creator economy" for artists, writers, podcasters, and musicians (De Cremer, Morini Branzino, and Falk 2023) and (on the very next day) that the technology will "change the nature of how we interact with all software" and "drive and distinguish how more brands compete" (Edelman and Abraham 2023). Broader still, *Foreign Affairs* prophecies a pivotal role for an "AI Economic Revolution" in reversing a millennial slowdown in productivity that, they argue, is destabilizing the global economy: "AI," write James Manyika and Michael Spence (2023), "holds the potential for a digitally enabled surge in productivity that could restore growth momentum" by easing the "shrinking labor pool in many countries." Yet, as the *MIT Technology Review* reported in January 2024 (Heaven 2024; Rotman 2023), more than a year into the commercial phase of this supposed revolution, there is still no certainty about the economic impact of AI and no rising prosperity for anyone outside the elite echelons of power.[12]

We have titled this special issue "Beyond Chatbot-K" in a nod to the harum-scarum tempo and nonstop barrage of new products, marketing claims, research findings, bombshell news stories, tech leader declarations, and business reports amid the relentless efforts of players large and small to vie for media attention, investment, regulatory advantage, and consumer uptake. As the journalist James Vincent (2023) writes, the tech industry for decades has propagated a "fallacy of version numbers" to convey the illusion that "numbered tech updates reflect definite and linear improvements in capability"—a strategy now recurring in generative AI. "Beyond Chatbot-K" is thus our way of suggesting that while we have *no idea* what version, product, or claim will dominate the news as readers encounter these words, we have organized this special issue to provide analyses, propositions, and thought experiments that, we hope, will retain insight and relevance well beyond the current version or

[12] See also Gardizy and Holmes 2024 for reporting on behind-the-scenes business discussions intended to "temper expectations" given that "hype about the technology has gotten ahead of what it can actually do for customers at a reasonable price."

wave.[13] The wide-ranging response to our call for essays and "thinkpieces" means that our special issue is divided into two successive parts. Hence, while this introductory essay surveys the contents of part 1—which include an opening essay on the technological transformations behind the rise of chatbots; a conversation with linguist Emily M. Bender and author Ted Chiang; and a diverse array of essays, shorter thinkpieces, and book reviews—it also alludes to a few of the essays that will appear in part 2 in the October 2024 issue of *Critical AI*.

The remainder of this introduction, both historical and analytic, accentuates the point that the now commonplace notion of commercial chatbots as the bleeding edge of an AI "revolution" is a relatively recent happenstance. As Matthew Stone, Lauren M. E. Goodlad, and Mark Sammons elaborate fully in their opening essay on this topic, the turn to chatbots was precipitated by self-supervising architectures like the generative pretrained transformer (GPT). Thus, the history of LLMs based on these architectures was not rooted in the design of interactive "chat" systems or in robust frameworks for humanlike communication or information access; the origins of LLMs, rather, were in the steady advance of statistical proxies for predicting the plausibility of automated transcriptions and translations. The rise of such powerful models for the probabilistic scoring of text meant that an increasing faith in the wonders of scale and the effectiveness of data attached to a technology that generates convincing output without any means of tracking its provenance or ensuring its veracity. As we will see, those little-discussed origins assure that even as state-of-the-art chatbots are touted as scientific marvels and ideal companions, their underlying function depends on the expropriation and privatization of human-generated content (much of it under copyright); the expenditure of enormous computing resources (including energy, water, and scarce materials); and the hidden exploitation of armies of human workers whose low-paid and high-stress labor makes "AI" seem more like human "intelligence" and communication.

That these material conditions of possibility are so hard to perceive is, as we will show, the result in part of misleading terminology, flawed technical benchmarks, and proprietary datsets—all of which make it easy to overstate the strengths of new generative products and difficult to verify real scientific advances. The use of anthropomorphic terms (*reasoning*, *learning*, *experience*, etc.) to describe what, in actuality, are mere quantitative proxies for the sociocultural and biological foundations of human understanding of, and communication about, a diverse world compounds what, according to Inioluwa Deborah Raji et al. (2021), is a crisis in *construct validity* (defined as the ability of a metric to measure or verify what is claimed). As weak benchmarks and tests purport to confirm the human-resembling capacities of machine intelligence, they obscure key differences (cf. Smith 2019; Newfield 2023).[14] Thus while

[13] For those unfamiliar with the usage, we note that *k* (comparable to *x* or *n*) is often used in mathematics and computer programming to signify a variable that serves as a placeholder for a constant value that has not yet been determined.

[14] See Goodlad 2023 for a discussion of the origins of "intelligence" in the field of so-called AI, and of the persistent tendency of some researchers to refuse clear definitions of machine intelligence as distinct from crude anthropomorphizing analogies, reductive proxies, and misleading benchmarks. It is worth

data-driven ML and DL are powerful technologies that can be selectively beneficial to human and planetary thriving (particularly when they leverage high-quality data and strong domain expertise), state-of-the-art machine intelligence (including generative AI) remains fundamentally narrow—a proposition that we substantiate in dialogue with the work of Stone, Goodlad, and Sammons in this special issue. It follows that educating the public in critical AI literacies, and democratizing conversations about AI design, implementation, and governance accordingly, are crucial conditions for the equitable and safe development of new technologies. At the broadest level, this coauthored introduction and the special issue that follows seek to mobilize the interdisciplinary resources necessary for advancing those timely projects.

The Prehistory of Chatbots

Although the term "AI" was coined in the 1950s, it has undergone multiple transformations as a field of research and, until recently, was familiar to the general public largely as a theme for science fiction. Readers of this special issue may have the impression that language models and chatbots have for decades been a kind of holy grail for computer scientists in academia and industry—but that is not actually the case. In fact, while they have long invoked chatbots as provocations and thought experiments for broader research practices, computer scientists have historically been quite skeptical of the practical utility of leveraging such systems as potential "AI" companions. For example, when the computer science pioneer Alan Turing (1950) proposed assessing the capabilities of intelligent machines by probing their abilities to respond in humanlike ways in text conversation, his work did not, as some might imagine, launch a race to build bots that could pass a so-called Turing test. Rather, Turing's work helped to crystallize the insight that a truly humanlike ("general") machine intelligence would need to be sufficiently versatile to communicate coherently in response to open-ended interrogation about a complex world. Thus, at least for some decades, attention to Turing's ideas encouraged an engineering mindset in academic research focused on careful evaluation of system behavior. Similarly, when philosopher John Searle (1980) offered a critique of mechanistic rule following as a strategy for simulating conversational interactions with computational systems—his famous "Chinese room" thought experiment—he proposed, in effect, that humanlike language understanding is a necessarily conscious phenomenon, inaccessible to machines and inseparable from people's situated encounters with one another and the world. An enduring lesson of Searle's thought experiment is that chatbots based on specific textual rules are extremely impoverished without perception and embodiment (see, e.g.,

recalling that the organizers of the "Dartmouth Summer Research Project" that coined the term "AI" envisioned a two-month, ten-person study of artificial intelligence that would "proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer" (McCarthy et al. 1955).

Harnad 1990; Bender and Koller 2020) and cannot feasibly achieve the open-ended interactivity of human responses (see, e.g., Shieber 2007).[15]

Although Joseph Weizenbaum's (1966) ELIZA program is now remembered as an early deployed chatbot with research impact, for many years it was a rare exception. ELIZA was a simulation of a Rogerian psychotherapist that Weizenbaum devised using simple pattern-matching techniques; in conversation it offered open-ended follow-up questions derived by rewriting user utterances (e.g., "What makes you think that . . . ?"). The lessons of this work were social rather than technical: Weizenbaum noted the problematic tendency of users to treat the program as if it were a human conversant in ways its simple functionality could not possibly justify. As he famously wrote, he had discovered that "extremely short exposures" to ELIZA's "parody" of psychotherapy "could induce powerful delusional thinking in quite normal people" (Weizenbaum 1976: 7). Such worrisome projections of human traits onto conversant software, which have come to be known as the "ELIZA effect" (Hofstadter 1995), predispose casual users to overestimate the capabilities of chatbots and to reflexively accept their outputs as if they were human-level and authoritative. For Weizenbaum, this tendency was so concerning that it called into question whether dialogue systems could ever be ethically deployed. In the decades after his intervention, academic researchers struggled to ensure that their models' capabilities were reliably assessed and judiciously applied as tools without amplifying the ELIZA effect. For example, in 1991, when philanthropist Hugh Loebner funded a chatbot competition modeled on the Turing test, the consensus in academia, as voiced by computer scientist Stuart Shieber (1994), was that asking judges to score conversations as humanlike is a deeply problematic process and that chatbot competitions could not and would not spur significant technical progress.

In the 1990s through the 2010s, researchers in the field of natural language processing (NLP) did not by and large pursue unstructured and open-ended approaches to interactive dialogue. Instead they tended to work toward highly regimented systems that could assist users in limited applications. As Stone, Goodlad, and Sammons explain in their essay, an important example was the Cognitive Assistant that Learns and Organizes (CALO) project led by SRI international from 2003 to 2008 and funded by the US Defense Advanced Research Projects Agency in the "Personalized Assistant that Learns" program. The project developed targeted methods that might allow future interactive computer systems to "reason, learn from experience, be told what to do, explain what they are doing, reflect on their experience, and respond robustly to surprise" and created software frameworks for integrating these prospective capabilities.[16]

[15] For a collection of critical and technical papers on verbal behavior as a hallmark of intelligence, see Stuart M. Shieber's (2004) excellent volume. As Katherine Bode and Lauren M. E. Goodlad (2023b) discuss at some length, Emily M. Bender and Alexander Koller's (2020) thought experiment about an octopus updates Searle's Chinese room for the DL era to argue that modern LLMs do not "understand" natural language in a humanlike way.

[16] "Artificial Intelligence: CALO," SRI International, no date, https://www.sri.com/hoi/artificial-intelligence-calo/. Accessed May 25, 2024.

Many readers will notice that even this regimented ML agenda is shot through with anthropomorphisms that suggest an unchastened determination to deliver something like human-level general intelligence ("AGI") through computational means. As Francis Hunger observes in a thinkpiece in part 2 of this special issue, anthropomorphizing metaphors (including the use of terms such as *learning*) are a serious impediment to public understanding of what modern chatbots are and do. This line of critical inquiry points to a little-discussed disconnect between, on the one hand, an ambitious discourse dating back to midcentury cybernetics (when "AI" was first inaugurated as a Cold War–era research enterprise and then formalized under the sway of "symbolic" approaches to simulating human cognition)[17] and, on the other hand, the implicit qualifications that researchers made as they shifted to task-specific goals in order to engineer piecemeal approaches to useful "cognitive assistance." In this practical context, when CALO researchers alluded to *reasoning*—a term that continues to prompt debate in AI contexts—they were not imputing humanlike rationality or generalizability to computational models. Rather, they were trying to break from a paradigm that required step-by-step coding for each and every task by envisioning more versatile tools. "Reasoning" from this perspective involves the ability to leverage NLP insights to execute focused tasks in context-sensitive ways—for example, scanning a user's emails, identifying a developing situation in doing so, and offering the option of scheduling a meeting in response. (It is worth clarifying that digital assistants that deliver this level of task-based "reasoning" do not yet exist.)[18]

Likewise, when computer scientists speak of machine learning, the *learning* in question denotes a computer model's ability to "optimize" for useful predictions during a process of "training" on data that involves the updating of numerical values in an elaborate set of statistical calculations. Thus, "learning from experience" does not invoke *experience* in, for example, either of the influential senses that the cultural theorist Raymond Williams (1985: 128) explored when he distinguished between two prevalent meanings of "experience" that have informed modern thinking since the Enlightenment (the first involving "past 'lessons'" and the second "full and active 'awareness'"). Rather, as if tacitly to devise a technological proxy for Williams's "past 'lessons'"—and with no explicit suppositions about "awareness" (full or otherwise)—the technologists of this period understood "experience" to consist in the potentially useful data (possibly quite different from the training data) obtained by the program after deployment ("in the wild"). From an ML standpoint, that is to say, "experience" equates to the acquisition of new data obtained in the wild, while "learning" from (or "reflection"

[17] For a thoroughgoing discussion of "symbolic" approaches to modeling humanlike cognition, sometimes called "GOFAI" (Good Old-Fashioned AI), see Smith 2019 and (for discussion of the book) Newfield 2023.

[18] As computer scientist Melanie Mitchell (2023) puts it, *reasoning* "is an umbrella term that includes abilities for deduction, induction, abduction, analogy, common sense, and other 'rational' or systematic methods for solving problems." The process of reasoning may involve "composing multiple steps of inference" that require "abstraction"—the capacity to apply learning developed from one example in relation to any number of examples. Though the lack of robust benchmarks obscures the question, today's LLMs do not reason in this "general" way.

on) such “experience” involves modes of statistical optimization informed by access to this new data during subsequent rounds of training or fine-tuning.

Some readers may regard this array of specialized meanings as a kind of rhetorical time bomb or slippery slope: a situation in which technologists wield ordinary language in irregular ways as they strive to conceptualize and implement computational proxies for complex human faculties that many people simply take for granted. While this use of commonsense vocabulary enlists human standards that might ultimately help technologists to situate the limitations of their computational proxies, until then it seems bound to proliferate terminological ambiguities and debates both inside and outside technology domains. Nonetheless, it is worth noting that one of the best-known commercial technologies of our time, Apple’s Siri assistant, was the product of this practical research paradigm. Originally launched on the iPhone 4S in 2011 (after Apple bought the start-up that SRI spun off to commercialize CALO technology), Siri possesses a fine-grained ability to control cell phone features in conjunction with simple (but highly useful) functionalities for web search and question answering. That the lingering anthropomorphisms of Cold War–era and symbolic “AI” could coexist with—and even help to usher in—practical projects that did not push an explicitly hyperbolic agenda is demonstrated by the fact that Siri-like digital assistants became popular features of everyday life without the companies that marketed them portraying these products as feats of (or pathways to) humanlike “AI.” With no sales pitch pressuring them to believe otherwise, users with little technological expertise could understand Siri as a tool whose delivery of information involves searching the web. Hence, despite the system’s signature ability to recognize spoken language and answer in return, the idea that Siri might become an intentional or superintelligent agent was virtually unheard of.

It was not until Amazon’s launch of the Alexa Prize socialbot competitions in 2016—that is, less than a decade ago—that chat interaction became a prominent domain of academic research (Khatri et al. 2018). Conceived to develop the potential for Amazon devices (such as the Echo smart speaker) to engage in “natural, sustained, coherent, and engaging open-domain dialogs” (Khatri et al 2018:1), this still ongoing competition (which completed its fifth round in September 2023) funds university teams to build bots that can interact via speech and screen.[19] The grand challenge of the prize—which no team has yet achieved—is for bots to “earn a composite score of 4.0 or higher (out of 5) from a panel of judges, and have those judges find that at least two-thirds of their conversations with the socialbot in the final round of judging remain coherent and engaging for at least 20 minutes” (Alexa Prize Team 2023). Clearly, by 2016 academics had abandoned their reluctance to legitimize and participate in “Turing test”–like endeavors. Indeed, Amazon’s largesse and its research ecosystem created unique opportunities for the academic research community—part of a wider trend in which some high-profile researchers work across industry and academic settings.

[19] Each team is supported by a substantial research gift—$250,000 in recent years—as well as technical infrastructure, computational services, and evaluation reports.

Three further trends were integral to building enthusiasm among researchers inside and outside academia for a new generation of chatbots. The first was a major financial incentive rooted in the business model that, in the 2000s, had turned tech companies such as Google (now Alphabet) and Facebook (now Meta) into surveillance empires: that is, into powerful purveyors of data and advertising which established them among the world's largest and most profitable corporations (see, e.g., Zuboff 2018; Doctorow 2020). From this standpoint, interactive chatbots designed to facilitate purchases, to enhance the presentation of paid content, and to expand user surveillance through a network of consumer-facing devices (in cars, homes, workplaces, "smart" cities, and so on) could become significant drivers of profit as they boosted the monetization of people's online activities.

The second, related trend involves the utility of data itself in improving the LLMs on which chatbots are built. These "deep" and fundamentally data-dependent statistical message-passing architectures improve considerably with the availability of ever larger data sets. From a purely technical outlook, such data-driven approaches thus offered efficient alternatives to the intensive purpose-built programming that had been required previously to assemble intelligent assistants through an assortment of preprogrammed skills. Socially and technopolitically, however, the growing belief in what an influential Google essay would call the "unreasonable effectiveness of data" (Halevy, Norvig, and Pereira 2009) produced a powerful discourse of *data positivism*, which continues to hold that the unprecedented scale and variety of available data constitutes a wholly new onto-epistemic substrate. As Katherine Bode and Lauren M. E. Goodlad (2023b) argued in their introduction to *Critical AI*'s "Data Worlds" special issue, data positivism narrows the framework for the production and evaluation of knowledge while legitimating a political economy that amplifies surveillance, the concentration of power, and the perpetuation of algorithmic discrimination. Although such data positivism predated the advent of large models on the scale of Google's BERT or OpenAI's GPT-2, LLMs (and their multimodal successors) are paradigm cases for uncritical assumptions about the onto-epistemic powers of data-derived pattern-finding. As this paradigm became ever more entrenched through the discourse of foundation models, industry researchers charged with preparing generative technologies for commercialization began to reconceive these systems in terms of a "misalignment" with human needs that could be gradually "aligned" through the massive enlistment of low-paid human labor—problematic techniques that Stone, Goodlad, and Sammons elaborate at length.[20]

The third trend concerns how, together with the growing availability of human-generated data on the internet and through networked devices, advances in computing helped to constitute "deep learning" (DL) as the "go-to" framework for modeling data at scale.[21] In 2007, manufacturers of graphical processing units released software that

[20] As Luke Stark (2023: 370) notes, "That technical solutions are often touted as the answer to AI's inadequacies by the same companies developing and profiting from these systems in the first place suggests AI alignment is the wrong way to think about the broader questions at stake."

[21] See also Malevé and Sluis 2023 for a discussion of the photographic pipeline for AlexNet, a landmark DL software architecture dating back to 2012. To be clear, what is "deep" in DL refers to the multiple

enabled researchers to substitute the latter for standard computer hardware (central processing units)—an innovation that fueled a massive speedup in the performance of the mathematical operations required for DL. At the same time, computer memory and data storage capacities dramatically increased. These enhanced hardware capabilities allowed researchers to create larger models and more complex architectures, making the scale required to apply DL techniques to language and vision problems feasible for the very first time (Malik 2021; Pandey et al. 2022; Thompson et al. 2020). By the 2010s—the decade in which AlphaGo, a DL system, defeated the world's best human Go player—the tech industry, having recognized the commercial potential of this resource-intensive paradigm, began to develop monetizable applications.

As Meredith Whittaker (2021: 51–52) has observed, with the growing enthusiasm for DL's commercial value, tech companies began to leverage the rhetoric of "AI" as a powerful "marketing hook" for such "data-dependent approaches." By portraying "AI" as "the product of breakthrough scientific innovation," they downplayed the importance of concentrated data and computational resources (cf. Kak and West 2023). Hitherto best-known to the public as a topic for movie screens and sci-fi classics, "AI" became synonymous with new or improved applications such as voice interfaces, translation tools, recommendation systems, and automated driver assistance. The stage was set for language models trained on data sets estimated to require twenty thousand years for a human to read (LeCun 2023) to come to the fore as "foundations" for what is sometimes adduced as a "fourth industrial revolution" (Bommasani et al. 2021; McKinsey 2022).[22]

Social, Cultural, Politico-economic, Technical, and Environmental Implications

Having unpacked the socio-technical history of generative AI and the political economy on which it rests, we now turn to the consequences of the widespread operationalization and commercialization of these new technologies. To be clear, none of these implications arose de novo in response to the commercial development of generative pretrained transformers. To the contrary, as we have already suggested, many seeds were planted decades ago with the monetization and capitalization of online data, the consequent rise of data surveillance as a lucrative business model, and the increasing concentration of power and resources that inevitably followed. To varying degrees, such conditions apply to all large-scale "AI" models developed commercially—not just OpenAI's products and not just LLMs, chatbots, or even

layers in a pretrained model through which new inputs pass before delivering one or more predictive outputs.

[22] According to a Stanford University report (Bommasani et al. 2021), a *foundation model* (a term Stanford researchers coined and have since popularized through their Center for Research on Foundation Models) "is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks." Though still controversial, the notion of LLMs as "foundations" for future research and applications, which Stanford thus established in 2021, reinforces the idea of resource-intensive and flawed LLMs (and multimodal successors), as the dominant paradigm in "AI" research even though these models are recognized as being "misaligned" with safe human use (see Ouyang et al. 2022 as well as Stone, Goodlad, and Sammons in this special issue).

generative AI. Nonetheless, increasing expectation of an "AI revolution" through what enthusiasts envision as an emerging trillion-dollar market for chatbots and other generative tools has amplified the overlapping social, cultural, politico-economic, technical, and environmental implications that we now elaborate.

Amplification of Bias and Stereotypes

As many researchers have documented, the tendency of some computational systems to amplify bias and stereotypes is endemic to a paradigm in which statistical patterns serve as questionable proxies for lived knowledge of (and in) a diverse world (see, e.g., Noble 2018; Benjamin 2019; Devinney, Björklund, and Björklund 2022; Broussard 2023). In their now canonical article on the dangers of LLMs, Bender et al. (2021: 615) show how such supersize models are subject to *documentation debt*, a situation in which the uncurated data sets used for training are "too large to document post hoc." Whereas "documentation allows for potential accountability," they explain, "undocumented training data perpetuates harm without recourse." Thus, according to one recent study, LLMs "exhibit archaic stereotypes about speakers" of African American English that resemble "the most negative ever experimentally recorded human stereotypes about African Americans, from before the civil rights movement" (Hofmann et al. 2024: 2). Moreover, because generative models reproduce such problematic content through synthetic outputs (Shah and Bender 2024), they strip it of an "original context" in which biases and stereotypes are "more clearly situated as the ideas of people." In doing so, LLMs and other generative systems endow prejudicial content with the false authority of machines.[23]

Misinformation and Degradation of the Internet (through Misconceptions, Conspiracy Theories, "Hallucinations," and Malicious Use)

Because generative models do not understand language in a humanlike way—despite their fluent modeling of linguistic forms and patterns observed in training sets—they cannot distinguish between truth and falsehood, or recognize inappropriate stereotypes and biases**.** The result is that LLMs and other generative models are likely contributing to the stream of socially and politically destabilizing misinformation on social media, dubious websites, and a degraded online ecosystem. According to one expert cited in

[23] Through probing and audits of LLMs, researchers have discovered "persistent toxic" content (Gehmen et al. 2020: 3356) and "severe" bias against Muslims (Abid, Farooqi, and Zou 2021: 298). On gender bias see, for example, Sheng et al. 2019; Lu et al. 2018; and Kotek, Dockum, and Sun 2023. On such biases with regard to synthetic letters of reference, see Wan et al. 2023. Looking at multimodal models, Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe (2021) found misogynistic and pornographic content and "malignant stereotypes"; Andrew Hundt et al. (2022: 753) warn that robots programmed with CLIP (an OpenAI image-to-text classifier) pick up stereotypes including "racist, sexist, and scientifically discredited physiognomic behavior"; and Federico Bianchi et al. (2022) have documented the amplification of demographic stereotypes in large image models (see also Drahl 2023). For documentation of dubious outputs that discuss the (alleged) benefits of slavery and fascism, see Piltch 2023; for a study of "trustworthiness" that shows the ease with which GPT models can be prompted to generate biased content, see Wang et al. 2023. See Omiye et al. 2023 for a study of how LLMs propagate "race-based medicine" and other "harmful ideas about race; and see Bai, Wang, Sucholutsky, and Griffiths (2024) for the use of prompt-based methods to detect and measure implicit bias in LLMs.

the *New York Times*, generative tools such as ChatGPT are powerful engines of misinformation that enable "false narratives" to be "frequently" spread "at dramatic scale" (quoted in Hsu and Thompson 2023; cf. Milmo and Farah 2023). Tech columnist Julia Angwin (2023) also worries that the internet will become "even more polluted with untrustworthy content." "While creators of quality content are contesting how their work is being used" and may therefore hesitate to post online, she explains, "dubious A.I.-generated content is stampeding into the public sphere."

LLMs are ideal for propounding false information for at least four technical reasons. First, the systems are prone to mimicking and, indeed, amplifying the human-generated misconceptions, falsehoods, and conspiracy theories that are rife in training data scraped from the web (Sharon Levy, Saxon, and Wang 2021; Lin, Hilton, and Evans 2022; Khatun and Brown 2023). Second, the models are systematically vulnerable to a type of inept prediction that the industry designates through the misleadingly anthropomorphic language of "hallucination" (cf. Klein 2023; Knibbs 2024). Also referred to as "confabulations," these defective outputs—defined by Ziwei Ji et al. (2023: 3) as generated text "that is nonsensical, or unfaithful to the provided source input"[24]—are inseparable from how current generative systems work. As Stone, Goodlad, and Sammons emphasize in this special issue, LLMs are designed to generate plausible replies to any and every prompt but not to verify the accuracy or suitability of these outputs. Research efforts of many kinds have failed to eradicate this core technical weakness: at present, even the task of identifying problematic outputs—much less of reducing their frequency—typically requires human intervention (Thomson and Reiter 2020; Ji et al. 2023).

For example, Abeba Birhane and Inioluwa Deborah Raji (2022) report that ChatGPT produced text on "how crushed porcelain added to breast milk can support the infant digestive system." A more commonplace example—as lawyers using chatbots for legal "research" have discovered to their chagrin (Weiser 2023; Associated Press 2023)—occurs when LLMs fabricate credible-seeming facts or citations. The training of LLMs is sometimes likened to a kind of "Mad Libs" regimen during which the system identifies common patterns that the model then leverages for responses to user prompts.[25] Hence, when coauthor Goodlad directed GPT-3.5 to complete a prompt that identified her as "a professor of English" (and so on), the system, in Mad Lib–like

[24] Though we concur with Ji and colleagues' definition of the defective outputs in question, we do not share their view that this problem corresponds to hallucination in humans. That is, while Ji and colleagues rightly identify *hallucination* as a "psychological term" pertinent to a "waking individual, in the absence of an appropriate stimulus from the extracorporeal world," or as "an unreal perception that feels real" (3), they do not explain how the "nonsensical" output of a disembodied statistical model corresponds to those conditions. Consider that LLMs are not "individuals," are not "corporeal," do not sleep or wake, do not have psychologies, do not "feel" reality (or anything else), and, as currently designed, receive no "stimulus" from the world (apart from inputs delivered by users). Ji and colleagues' language is thus shot through with the kind of reflexive and unsubstantiated anthropomorphisms that this special issue questions.

[25] See also Chiang's (2023) description of ChatGPT as a "blurry JPEG" of its training data—a case of "lossy compression."

fashion, generated information that—though wholly false—resembled content found on many faculty web pages (see fig. 1).[26]

<COMP: Please place figure 1 about here.>

A third reason for machine-generated misinformation concerns a lack of access to information that postdates the training data: current attempts to mitigate that problem through hybrid systems that combine chatbots (such as Microsoft's Copilot or Alphabet's Gemini) with search engines (such as Bing or Google) are, at least for now, unreliable.[27] Hence, when Goodlad used a similar prompt to ask Bing Chat (now Copilot),which harnesses GPT-4, to write a "bio" for her, the combined system successfully synthesized some of the content from her current Rutgers web page (see fig. 2). **{Au: Please confirm change to "2" Confirmed}** However, by describing her as a professor at the University of Illinois, the output reveals that the system's locating of an active faculty web page at Rutgers did not result in Bing Chat's correction of a prompt with out-of-date information. Moreover, when Bing Chat attempted to synthesize information from the active web page, it falsely reported that Goodlad's PhD is from Rutgers (where she has taught since 2017) and that she taught at Rutgers until 2017 (whereas in fact she taught at Illinois until 2017).[28] Such unreliable outputs confound efforts to use LLMs for information access (see also Allison and DeRewal in part 2 of this special issue); but as Stone, Goodlad, and Sammons elucidate, these results are entirely consonant with the historical rationale of LLMs: that is, plausible texts should be scored highly by systems whose overarching goal is to resolve ambiguities in systems for machine transcription and translation.

<COMP: Please place figure 2 about here.>

It follows that, as Chirag Shah and Emily M. Bender (2024) remark, simply because a system has modeled a huge trove of information "does not make it knowledgeable or capable of producing reliable answers." Moreover, since the sources the LLM has trained on are buried in huge, proprietary, and undocumented data sets, such tools make it impossible for users to check the precise provenance of any given output (cf. van Rooij 2022). Chatbots, as Shah and Bender conclude, are "beneficial" primarily when the content of language is unimportant.[29] That is why Leslie Allison and

---

[26] In this initial prompt, Goodlad identified herself as a professor at the University of Illinois (where she taught between 2001 and 2017) in the effort to aid the model, which, she knew, had no access to training data after 2021. Goodlad thanks Anna Mills for her assistance with this probing experiment.

[27] On Google's renaming of the Bard chatbot as Gemini (already the name of the latest underlying model), see, for example, Goode 2024. On the use of a technique called retrieval augmented generation (RAG) in designing such systems, see also Lewis et al. 2020 and note 51 below.

[28] Whereas Goodlad earlier had identified herself as an Illinois professor in the effort to maximize GPT-3.5's pre-2021 training, in this case she did so to test Bing Chat's ability to contradict an erroneous prompt.

[29] As an example of a potentially beneficial usage, Shah and Bender (2024) seize on chatbot translation of one form of computer code to another—a type of machine translation that can be checked for accuracy by compiling and running the code. However, "beneficial" uses, they argue, must be evaluated alongside outstanding questions of harm.

Tiffany DeRewal, in their contribution to part 2 of this special issue, argue that "AI" is making it harder for people to "find trustworthy sources" and, just as important, "to *know*" when they have done so.[30]

In his thinkpiece in this special issue, literary scholar Aaron Hanlon delves into the problem of generative AI and veracity by trying out the idea that the outputs of LLMs can be treated as nonliterary fictions. The tendency of chatbots, for example, to fabricate quotations from W. E. B. Du Bois's *The Souls of Black Folks* is a dilemma, according to Hanlon, exacerbated by "mismatched expectations." Drawing on Hannah Kim's work on the philosophy of fiction, Hanlon compares and contrasts a chatbot's utterances to those of characters in a play and imagines the potential for human users to accustom themselves to regarding bots through a fictional lens. As against the tendency of developers to encourage users to regard chatbots as if they were search engines or humanlike companions, people must learn to treat these systems "provisionally" and evaluate their content accordingly. Hanlon's thinkpiece provides an interesting comparison to the arguments of African American studies scholar Maurice Wallace and undergraduate student Matthew Peeler. According to their coauthored thinkpiece, Khan Academy's recent turn to "animating" figures like Harriet Tubman exemplifies a misguided "messianic faith" in AI's alleged ability to "resurrect the dead" for the supposed "salvation" of engaged historical learning. Despite deceptive claims of "bringing history back to life," the Tubman chatbot is "little more than an advertisement for Khan Academy (and its ostensibly liberal politics of race)"; the "whitewashed" chatter of "AI Tubman" is closer to a Wikipedia entry than the speech of a nineteenth-century freedom fighter.[31] From this view, Hanlon's case for the fundamental fictionality of text generation runs up against the fundamentalism of developers who neither admit the limits of probabilistic mimicry nor recognize the "minstrel ridicule" implicit in marketing bland platitudes as the language of the historical Tubman.

That such "minstrel ridicule" might sometimes be deliberate takes us to the fourth technical source of misinformation: the ease with which malicious users can circumvent the weak guardrails put in place to ward against the generation of harmful content. The practice of "jailbreaking" ChatGPT (see, e.g., Christian 2023) has become a comical pastime, but malicious use of generative AI, which may include deepfakes, political misinformation, and nonconsensual pornography, is, of course, no laughing matter.[32]

---

[30] On Google's controversial incorporation of "AI Overview" as the default for its signature search engine in May 2026, resulting in false outputs such as the mistaken identification of Barack Obama as the first Muslim president of the United States, see Grant 2024.

[31] On the subject of Khan Academy founder Sal Khan's long-standing criticism of education (often proffered in collaboration with Bill Gates), Audrey Watters (2021: 5) writes that Khan's claim that US education has been static since 1892 is "woefully inaccurate—offensively so."

[32] On the potential dangers of AI-generated disinformation from foreign adversaries, see Byman et al. 2023. On the surging "deepfake" porn economy, see Tenbarge 2023; and see Maiberg 2023 for a disturbing account of how generative models are used to "produce any kind of pornographic scenario . . . trained on real images of real people scraped without consent from every corner of the internet." See Funk, Shahbaz, and Vesteinsson 2023 for a report documenting how generative tools are being used to "supercharge online disinformation campaigns" and to "strengthen censorship" in authoritarian countries.

Copyright Infringement, Lack of Consent

The use of copyrighted content scraped from the web without consent for the training of generative models opens a host of legal questions—many of which exceed the topic of LLMs per se (see, e.g., Appel, Neelbauer, and Schweidel 2023) and may be of special concern to performers and visual artists (see, e.g., Center for Artistic Inquiry 2023; Davis 2023; Jiang et al. 2023; Shepard 2024). In August 2023, the *New York Times* updated its terms of service to forbid use of its contents for training models (see Weatherbed 2023). In December 2023, the same newspaper sued OpenAI and Microsoft on the grounds that "millions of articles . . . were used to train automated chatbots that now compete with the news outlet as a source of reliable information" (Grynbaum and Mac 2023). The *Atlantic* documented that hundreds of thousands of copyrighted works are "secretly" being used to train proprietary models (Reisner 2023).[33] Notably, one point the industry's competing leaders appear to agree on is that the use of copyrighted content necessary to develop their models is not something they plan to pay for.[34]

That said, copyright is a narrow lens from which to view the lack of consent in AI training—a problem that also involves public resistance to the datafication of everyday communication and activities as well as society's prerogative to embargo illegal material. For example, a recent study (Cole 2023), comparable to earlier research by Birhane, Prabhu, and Kahembwe (2021), documented the use of illegal materials (including thousands of externally validated images of child sexual abuse) in the LAION-5B data set—a common source for the training of commercial image models.

In her article in this special issue, legal scholar Sylvie Delacroix steps back from these surveillant, extractive, and exploitative data practices. Her visionary alternative borrows ideas from ecocriticism and water rights to imagine a revamped legal framework for a "data trust." Defining data as "an intangible, nonrival good," Delacroix compares the corporate hoarding of data to the hoarding of water. She argues that current legal tools are insufficient to constitute and sustain the "data rivers" that can enable a socially sustainable data ecosystem to flourish.

Platforms that fail to edit properly are complicit in this phenomenon: for example, according to CNN, since Microsoft automated the editing of its influential news site, the software giant has published improper content (including a fake story that Joe Biden "fell asleep during a moment of silence for victims of the Maui wildfire," a claim that the latest COVID-19 surge was "orchestrated by the Democratic Party," and an obituary for an NBA player that described the player as "useless") (O'Sullivan and Gordon 2023). See Schiff, Schiff, and Bueno 2022 for discussion of public figures using disingenuous accusations of "fake news" to evade accountability for unflattering but truthful coverage; on recent systems for the generation of synthetic videos, an obvious resource for malicious actors, see, e.g., Cai 2023.

[33] See OpenAI 2024 for the company's response, and, for critique of the latter see Tangermann 2024a. See Knibbs 2024 for the growing problem of "scammy" AI-generated books that mimic copyrighted titles. For recent deals struck with the *Atlantic* and several other news organs, see De Vynck 2024.

[34] For example, according to *Business Insider*, tech companies including Meta, Google, OpenAI, and Microsoft claimed that paying for copyrighted material "would create an impossible hurdle," while the venture capital firm Andreessen Horowitz argued that "the billions of dollars" already pumped into AI "should be reason enough not to create any new rules to benefit copyright holders" (Hays 2023).

Surveillance and Privacy Concerns

The extraction of users' data as they interact with generative tools extends the regimes of surveillance that began with the monetization of social media and search engines. Generative AI thus exacerbates already pressing concerns over surveillance and data privacy (see, e.g., EPIC, n.d.).

According to Pratyusha Ria Kalluri et al. (2023), AI research (especially involving computer vision) has encouraged a deep culture of surveillance and data extraction that has permeated the elite universities and tech companies that conduct such research. Ajay Sudhir Bale et al. (2024) have explored "the ease with which" generative systems may leverage surveillance to produce "accurate and personalized material" and—more concerningly—"unauthorized content synthesis" that involves "false information, counterfeit goods, and other forms of illegal manipulation."[35]

In a climate still prone to "moving fast and breaking things," the consultants, journalists, and educators now arguing for (or simply anticipating) the widespread adoption of commercial bots in public education seldom consider the impact of such adoption on student choice and privacy.[36] Writing with both students and educators in mind, Kathryn Conrad's thinkpiece in this special issue, modeled partly on the Biden administration's call for robust protections "from abusive data practices" and "agency over how data . . . is used" (US Office of Science and Technology Policy 2022), specifies, for example, that students "should be able to opt out of assignments that may put [their] own creative work at risk for data surveillance and use." **{Au: I think the citation of the executive order was incorrect here—that order is discussed in Goodman's piece in this issue but is neither cited nor mentioned in Conrad's piece, which instead cites and draws from the Biden administration's "Blueprint for an AI Bill of Rights" (which is not an executive order). Since the first two of these quoted passages appear in the "Blueprint," I have added it to the ref list of the intro and cited it here.}** Educational institutions, Conrad emphasizes, "have an obligation to protect students from privacy breaches and exploitation."

Environmental Footprint

Amid a climate crisis that calls for "rapid, deep and in most cases immediate greenhouse gas emission reductions in all sectors" (IPCC 2023: 20), policymakers and

---

[35]The belief expressed by Bale et al. (2024) that "developers of technology may successfully handle problems and moral dilemmas while also driving the accountable and constructive growth of generative AI if they promote an environment of cooperation, information sharing, and moral accountability" overlooks a strong pattern in the contemporary tech industry wherein once corporations seek to monetize their products, and especially once they answer to investors, their commitments to "moral accountability," if any, become to some degree merely performative. In the face of competition for market dominance, "cooperation," and "information sharing" likewise decline. For interdisciplinary studies of data that zero in on data extraction and surveillance, see for example, Gitelman 2013; Sadowski 2019; D'Ignazio and Klein 2020; Denton et al. 2021; Chun 2021; Franklin 2023; Pasquinelli 2023; Bode and Goodlad 2023a.

[36] See Watters 2023 for an invaluable history that documents the origins of today's most boosterish and AI-friendly discourses of educational technology in repeated efforts to introduce "teaching machines" that date back to the 1920s.

the public at large must pay close attention to any activity that consumes eye-watering resources—and is growing rapidly.[37] Yet tech companies view the environmental footprint of their AI efforts as trade secrets, as vital to their competitive advantage as their data sets, software, and models. They also like to claim that supposed advances human-level AI technologies that do not (and may not ever) exist "will supercharge society's ability to tackle and manage climate change" (quoted in Climate Action against Disinformation Coalition 2024). As outside investigators offer estimates of the overall footprint of AI, they find that the demands of training new models are increasingly dwarfed by the impacts of running these models for broader commercial use. For example, a recent study of the carbon emissions required to build and train BLOOM (Luccioni et al. 2022)—an open-source model comparable in size to GPT-3—suggests that the latter model's training probably emitted 50.5 metric tons (approximately the total emissions necessary for driving an average car for 125,000 miles [Tso 2023]).[38] By contrast, as reported by *Business Insider* (Mok 2023), the "massive" computing necessary to perform the calculations required to answer simple user prompts may mean that the costs of operating the current version of ChatGPT exceed the costs of training GPT-3 "on a weekly basis."[39] According to *Bloomberg News* (Farhat 2024), the International Energy Agency reported in January 2024 that global demand for electricity from "data centers, cryptocurrencies, and artificial intelligence could double over the next three years, adding the equivalent of Germany's entire power needs." According to the *Atlantic*, the $10 billion that Microsoft is funneling into data center expansion every quarter marks what one analyst describes as "the largest infrastructure buildout that humanity has ever seen" (Hao 2024; see also Rathi and Bass 2024).

Coinciding with these intensive energy costs is the correlative usage of water to cool the servers that perform these computations. As Microsoft and Google ramped up the production of generative AI between 2021 and 2022, the former reported a 34 percent increase in its consumption of water (Mercado 2023; O'Brien, Fingerhut, and

---

[37] For example, numerous sources (e.g., *CIO Coverage*, n.d.) have estimated the computational costs of running ChatGPT at about $100,000 per day—but in the months since its release that figure has steadily climbed closer to $1 million per day (see, e.g., Gardizy and Ma 2023). See Leffer 2023 for an interview in which data scientist Alex De Vries proposes mandatory disclosures. In an interview in which his comments on energy requirements were offered in the context of Google's business model for so-called generative AI, CEO Pichai explained that the company was already projecting the growing costs of operationalizing large models twenty-five years out (Wheatley 2024). On OpenAI CEO Altman's headline-making call for energy "breakthroughs" to power AI development, see, for example, Tangermann 2024b.

[38] On the challenges of estimating the carbon footprint of particular models, see also Patterson et al. 2021. Strubell, Ganesh, and McCallum (2019) estimated the carbon emissions of training much smaller LLMs such as the 1.5 billion parameter GPT-2 at roughly equivalent to "nearly five times the lifetime emissions of the average American car," including the manufacture of the car (Hao 2019). On water usage, see also Li et al. 2023; Hao 2024.

[39] See also Weidinger et al. 2021: 37, where the authors argue that researchers should expect "companies offering services that rely on such models" to "spend more energy, money and time on operating such models than on training them." Lauren Goode 2024 speculates Google pivoted to a subscription model for its top-tier "AI" in order to help "defray the massive computing costs associated with training and running a large language model.

Associated Press 2023) while the latter reported a 20 percent increase (Langley 2023). This evidence for the growing environmental footprint for developing and deploying generative AI compounds the already hefty (but little-publicized) use of electricity, water, air, heat, metals, and rare earth minerals necessary to build and maintain the massive server farms on which many of today's computational systems run (Monserrate 2022).[40] Indeed, although what the media likes to call an AI "arms race" is often discussed as if growth in subscriptions were the primary commercial goal, the expanded demand for "cloud" services that generative systems require may be an even more attractive prospect for a company such as Microsoft.[41] Hence, as Mél Hogan argues in "The Fumes of AI," a thinkpiece in this special issue, exposing the "environmental costs and impacts" of generative technology to the public has become a crucial dimension of developing the critical AI literacies of legislators and the world at large.

Concentration of Resources

In March 2023, OpenAI released GPT-4, which, at an estimated 1.76 trillion parameters, is about ten times the size of its precursor and, according to CEO Altman, cost over $100 million to train (Knight 2023a). GPT-4—updated and marketed as "GPT-4 Turbo" (Edwards 2023b) and, more recently, as GPT-4 Omni (Open AI 2024b)—remains at the time of this writing the underlying model in the paid subscription to ChatGPT. As there is already talk of GPT-5, as well as new versions and products from numerous other companies, it is worth emphasizing that all of the major players involved in generative AI fine-tune and/or train new models at this resource-intensive scale more or less constantly.

As start-up founder Nasrin Mostafazedeh observes (in a thinkpiece interview in part 2 of this special issue), the formidable cost of training large models is beyond the reach of all but the richest technology companies. Not only must start-ups like Mostafazedeh's purchase high-priced computational services from the same corporations that market the largest models, but they may also need subscriptions to the models themselves. Companies that lack significant investment from large tech companies thus operate at a considerable disadvantage and face high barriers to entry (see also Wiggins 2024). The effects of this worrying concentration of resources include the capture of academic research (Whittaker 2021) and, as Mostafazedeh emphasizes, threats to innovation itself.

[40] According to one report (Shift Project 2019), digital technologies were emitting 4 percent of greenhouse gas emissions—more than civil aviation—out of which an estimated 19 percent was for server farms.

[41] As Paris Marx (2023) writes, even if generative AI is a "flash in the pan," Microsoft's partnership with OpenAI enabled it to "drive customers to its Azure cloud platform" and away from competitors like Amazon Web Services (AWS) and Google. Anecdotally, some developers that have the choice to work with either Microsoft's or OpenAI's version of GPT-4 have speculated that Microsoft's servers (which host both company's versions) privilege the computational needs of Microsoft's own subscribers. In other words, OpenAI and Microsoft are rivals as well as partners.

Ironically, the strongest evidence for this excessive concentration of power may be the case of Suleyman (whose promotional book tour and start-up in "empathetic AI" we describe above). Despite his ability to leverage the reputation of a co-founder of DeepMind (a prestigious Google subsidiary since 2014), and despite raising more than a billion dollars for his start-up Inflection AI (founded in 2022 in partnership with LinkedIn co-founder Reid Hoffman and with backers including Bill Gates, ex-Google CEO Eric Schmidt, and the chip giant Nvidia [Mann 2024]), Suleyman did not gain sufficient traction in the "race" to build a consumer-facing chatbot—a field dominated largely by Open AI (in partnership with Microsoft), Alphabet, Meta, and Anthropic (a start-up originally split off from OpenAI and whose multi-billion dollar investors include Amazon and Alphabet). Thus, in a move that surprised some while confirming the suspicions of others, Suleyman abandoned his unicorn start-up and its signature chatbot, and—along with most of his development team—joined Microsoft. As one expert cited in *Forbes* put it, "for Inflection to exit the consumer race so quickly—after having raised so much money, so fast—while its CEO departs for a top job at Microsoft, raised some eyebrows… [but] didn't fully surprise… The costs associated with the consumer race mean that only a small handful will survive and win" "It's a good day for [Suleyman] and a bad day for Inflection's investors and early employees" (quoted in Konrad 2024). As another quoted expert put it, "Big Tech" companies such as Microsoft are already the "expected" winners of the "consumer AI horse race" (quoted in Konrad 2024).

Proprietary Research/Hyperbolic Claims about "AGI"

Already implicit in Mostafazedeh's concerns, is how the concentration of resources molds the research environment in and through which "AI" is imagined, developed, deployed, and articulated to the public. As design justice theorist Sasha Costanza-Chock (2020: 6) observes, even as design mediates lived realities and exerts "tremendous impact" on social worlds, very few people—least of all those "most adversely affected by design decisions"—have the chance to "participate in design processes." **{Au: Please add this source to the ref list}** The political economy and climate under which generative AI now develops has taken this dilemma to a new extreme.

In their contribution to this special issue, Nia Judelson and Margaret Dryden argue that the industry's habitual lack of transparency is paradoxically covered over by the obscurantist rhetoric of DL technology. In particular, the metaphor of the *black box*—a term derived from midcentury cybernetics to describe opaque systems and now frequently applied to DL architectures—has become an impediment to public discussion and ethical deployment of LLMs and other generative tools by implying that "AI" is beyond the reach of oversight on purely technical grounds. In concert with Bruno Latour's (1999) observations about technological opacity and Bender et al.'s (2021) critique of undocumented data sets, the black box metaphor, according to Judelson and Dryden, "normalize[s] the assumption that complexity is opaque" and, in doing so, justifies "uncritical engagement and developer unaccountability at every part of the [LLM] production process."

Indeed, even expert researchers seeking to evaluate the behaviors of supposedly state-of-the-art commercial products must contend with lack of access to the relevant data sets, architectures, and algorithms. Tech companies and their supporters argue that this departure from the norms of scientific research is economically necessary, just as they claim that strong regulation or public interest guidelines would inhibit progress (as they define it). As technology journalist Karen Hao (2023) explains, scientific knowledge-sharing and consensus building becomes ever more vexed when research "once largely performed in the open . . . happens in secrecy." Without the opportunity for independent corroboration, a team's claims to a scientific "*breakthrough*" are at best provisional and, at worst, promotional: a claim "assigned by a small group of employees as a matter of their own opinion."

This culture of unverifiable claims combines with specialist vocabularies and flawed benchmarks to blur the lines between science and hype. AI research has long been plagued by vague definitions of "intelligence" and a habit of anthropomorphizing tools (see the thinkpiece by Hunger in part 2 of this special issue), but in the ChatGPT era, the casual use of "AGI" has become flagrant. As one tech reporter puts it, the "craze" for generative AI is fueled by "belief that the tech industry is on a path" to realizing "god-like intelligence" (Heath 2024). Thus "AGI," hitherto a loose acronym for the speculative goal of modeling humanlike reasoning, is now frequently bandied as if its achievement were imminent—as, for instance, in Altman's (2023) claims that OpenAI is "carefully steward[ing] AGI into existence." As if the term "foundation model" were not sufficiently suggestive of innovation, moreover, in July 2023, OpenAI (2023a) announced a group devoted to discussions over *frontier models*: "As part of our mission of building safe AGI," they declared, "we take seriously the full spectrum of safety risks . . . from the systems we have today to the furthest reaches of superintelligence" (OpenAI 2023b).[42] According to Meta CEO Mark Zuckerberg, the arrival of AGI will be gradual:

You can quibble about if general intelligence is akin to human level intelligence, or is it like human-plus, or is it some far-future super intelligence. But to me, the important part is actually the breadth of it, which is that intelligence has all these different capabilities where you have to be able to reason and have intuition. . . . I'm not actually that sure that some specific threshold will feel that profound. (Heath 2024) **{Charles: Style as extract, and the following as paragraph continued}**

Hazy though it is, Zuckerberg's evocation of diverse humanlike capacities ("to reason and have intuition") illustrates the point we made earlier—that a decades-long habit of loosely equating technical definitions of machine "reason" with human correlatives has contributed to a culture of unscientific and self-interested claims.[43]

---

[42] The same website (Open AI 2023a) asserts that "frontier AI models, which will exceed the capabilities currently present in the most advanced existing models, have the potential to benefit all of humanity." However, since "frontier AI" also poses "increasingly severe" and "catastrophic risks," the company declares its dedication to managing those risks.

[43] Altman's most recent claims on AGI, as reported by MacKenzie Sigalos and Ryan Browne (2024), are even more strikingly inconsistent. Stepping away from his own past flirtations with doomerism, Altman argues both that "AGI"—defined as the ability to perform tasks at human-level or above—is near at hand

There is, to be sure, some genuine ambiguity in the questions of "generalization" that the "G" in "AGI" evokes. When a large language model correctly infers a probable pattern from a similar pattern observed in training data, researchers may reasonably speak of "generalization." Nonetheless, such statistical inference anchored in large-scale pattern-finding does not in itself meet the bar for the humanlike "generalizing" that AGI has historically signified. Though data-driven DL at scale can be powerful and scientifically useful in diverse domains—oceanography, geometry, climate science, pharmacology, demography, and materials science among them—the technology itself remains fundamentally narrow (Pearl and Mackenzie 2019; Marcus and Davis 2019). Thus, despite the ELIZA-esque propensity of many people to believe otherwise, today's chatbots are likewise narrow: their modeling of linguistic forms, prediction of lexico-syntactic completions, and mimicking of training data does not actualize or simulate human-level hypothetical thinking, judgment, or metacognitive reflection, nor does it constitute evidence for humanlike experience or world models (Smith 2019; Bender and Koller 2020).[44]

In part 2 of this special issue, Bowman and two respondents (Matthew Stone and Wendy H. Wong) will explore the question of LLM capabilities at greater length. Bowman's essay surveys the arguments of those who hold that today's LLMs have developed qualitatively new capabilities—that is, capabilities that cannot be fully explained in terms of the narrow mechanics of training and optimization. He maintains that the question of how to contend with these "disruptive" tools requires reliable information about "technical capabilities, limitations, and trends" as well as "uncertainties that surround the technology." While we agree that good information is essential (and do not reject the possibility of "new capabilities" out of hand), we note that the greatest obstacle to information at present is the industry's own secrecy. To that degree, unaccountable self-promotion, spun in the media, has to some degree filled the vacuum where shared knowledge and transparency ought to be.

Consider, for example, "Sparks of Artificial General Intelligence" (Bubeck et al. 2023), a self-published paper from Microsoft that appeared shortly before the release of OpenAI's GPT-4, which attributes "new capabilities" to that model that cannot be explained in purely statistical terms. In contrast to Bender et al.'s (2021) characterization of LLMs as "stochastic parrots," Bubeck and colleagues argue that

(achievable in "the reasonably close-ish future"), and that "it will change the world" and "jobs much less than we all think." Of course, a bona fide ability to engineer the human-level or suprahuman performance of "tasks" (reminiscent of science fiction) would be extraordinary. Likely Altman has in mind the narrow range of tasks that LLMs now imperfectly perform (even as they rely upon human reinforcement to achieve that limited performance capacity). As if to have his cake and eat it too, Altman wants to designate this limited and human-dependent assemblage of generative tasks, as if it were on the cusp of something called "AGI," *and* to describe the impact of that attainment on the human workforce as moderate (Altman 2023).

[44] Another flawed argument for AGI's supposed imminence derives from the ability of the largest LLMs not only to generate text but also to generate code, translations, and some low-level mathematical operations. This ability to perform multiple tasks for which a system has trained on relevant data, as Goodlad (2023) has argued, "is now misleadingly conflated with 'general' intelligence in the human sense."

(despite their admittedly "subjective and informal" approach) they have perceived a "breadth and depth" of capabilities that "could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system."[45] Notably, the version of GPT-4 that underwrites that claim is not publicly available, while the usual secrecy around data sets and architectures remains firmly in place. Thus, in the face of high-profile avowals of an embryonic "AGI," independent researchers are left to make educated guesses.

Take the case of multiplication, which is among those "new capabilities" adduced to exemplify an allegedly imminent "AGI." We can infer, for example, that when GPT-4 answers three-digit multiplication questions correctly but fails to answer four-digit multiplication questions, the likelihood is great that the system's parameter space and training data included the resources necessary to answer three-digit (but not four-digit) questions. Thus, while multiplication technically counts as a "new" capability, the means through which GPT-4 acquired that performance capacity is unlike that of a calculator (a special-purpose device that is systematically programmed to execute a wide range of multiplication problems) and unlike that of a person (whose ability to multiply is usually learned through practice). When humans learn to multiply, they typically acquire skills that enable them to shift from three- to four- (or even five- or six-digit multiplication) with relative ease. By contrast, GPT-4's "new capability" is limited to correct inferences about a limited set of multiplication problems, of which a significant fraction were likely observed during training. Properly speaking, the system is not actually multiplying at all but rather providing answers to multiplication problems.[46]

Exploitation of Human Labor

For LLMs to be commercially deployed as chatbots, developers must find ways to reduce the frequency of confabulations, misinformation, stereotyping, toxicity, and misuse. OpenAI's own researchers recognize that such problems occur because of the "misalignment" between a technology designed to predict lexico-syntactic completions and the very different objective of following a "user's instructions helpfully and safely" (Ouyang et al. 2022). As Stone, Goodlad, and Sammons explain in this special issue, the company's approach to "aligning" LLMs centers on reinforcement learning from human feedback (RLHF)—a technique rooted in the decades-old industry practice of hiring human crowdworkers to improve the functionality of ML and DL systems (see, e.g., Ross et al. 2010; Irani 2013; Gray and Suri 2019; Crawford 2021, chap. 2). Describing the origins of this approach in microwork crowdsourcing platforms like Amazon Mechanical Turk, communications scholar Lily Irani (2013: 723) calls it "artificial

[45] Critics such as Gary F. Marcus (2023) hastened to charge that Microsoft had "put out a press release . . . masquerading as science" in advance of GPT-4's public release. See also Margaret Mitchell 2023 for a helpful series of tweets.

[46] See also Dziri et al. 2023: 9, for an essay suggesting "that the strong performance of transformers" on complex tasks such as multidigit multiplication "should be taken with a certain grain of salt" since the "desired solutions could be readily derived from input-output sequences present in the training data, allowing for shortcut pattern-matching." Likewise, Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo (2023) argue that supposedly "emergent" capabilities are "mirages" that are contingent on the researcher's use of particular metrics.

artificial intelligence"—that is, the use of low-paid, platform-mediated, "on-demand" workers to meet the shortfalls of "AI." As *Time* reporter Billy Perrigo (2023b) documented, the process of detoxifying GPT-3 involved the outsourcing of labor to workers in Kenya who earned less than $2 per day to label graphic descriptions of "child sexual abuse, bestiality, murder, suicide, torture, self harm, and incest." Hence, what high-profile developers publicize as the automation of human-level tasks and the imminence of "AGI" quietly relies on a vast and expanding human "underclass" (Dzieza 2023; Wong 2023).[47] In this way, developers of commercial chatbots leverage human labor to court the ELIZA effect, masking the defects of a problematic tool, while encouraging users to regard these systems as if they were trustworthy sources of knowledge and humanlike companions. In addition, as Elizabeth Losh explains in her thinkpiece in part 2 of this special issue, tech companies shift the burdens of their faulty systems to others (including educators, librarians, students, and the parents of schoolchildren).

Beyond Chatbot-K: Against Data Positivism and the Myth of Frictionless Knowing

In their contribution to this special issue, Stone, Goodlad, and Sammons delineate a history of interactive digital systems that crosses from Vannevar Bush's visionary "memex" at the dawn of the National Science Foundation, and the CALO era of practical engineering which culminated in the development of Siri, to the surveillance-driven, for-profit data positivism that, in recent years, has come to regard human communication as a problematic inefficiency in need of lucrative top-down solutions. Nonetheless, it is worth remarking that while tech leaders talk up these Promethean aspirations—and investors imagine once-in-a-lifetime economic bonanzas—the actual functionalities of generative AI are, at least for now, remarkably humdrum.

According to a typical website (Upwork Team 2023), today's state-of-the-art chatbots can automate "content generation," bring "creativity" into line with "marketing trends," and "create personalized [customer] experiences." Writing in part 2, Losh, who is a scholar of Rhetoric, describes ChatGPT's outputs as "simplistic vagaries" that lack the ability to enter into complex questions or to use evidence effectively. "As many faculty and doubtless some students have begun to recognize," she adds, "generated text…typically lacks a strong ability to engage with the meaning-making activities of real-world audiences, contexualize occasions for writing that involve challenging the status quo, and support controversial hypotheses that demand cutting-edge analysis and reading against the grain."

Early evidence suggests that chatbots can be useful for generating code, though to what degree, given high error rates, remains subject to debate.[48] If there is an

[47] The development of automated driving technology also relies on humans to label datasets, but in November 2023, several outlets (see, e.g., Kolodny 2023) reported that Cruise robotaxis also required periodic help from a human workforce of "remote assistants" to help with "tricky drives."

[48] See Kabir et al. 2024 for a study that found ChatGPT's responses to queries about software engineering to contain inaccuracies 52 percent of the time and to be "verbose" 77 percent of the time.

aspirational vision implicit in such technology—beyond some enhanced productivity (and reduced employment) for "content" workers and programmers—it must surely lie elsewhere in the ever-nebulous and hype-laden prospectus for what AI's "revolution" might deliver. Of course, many entrepreneurs augur (in Suleyman's words) a new age of "radical abundance" that will endow the world's people with "broadly equal access to intelligence"—in other words, the same wishful thinking that underwrote technolibertarian hyperbole about the World Wide Web thirty years ago. But in the actually existing world, the automation of low-level "content generation" is hardly utopian. Its most immediate effects are to diminish economic opportunities for educated workers, enlarge the global "underclass" of crowdworkers, commandeer scarce resources, exacerbate online "enshittification" (Naughton 2023; Doctorow 2024), and leave educators to struggle to preserve student learning (Goodlad and Conrad 2024). "Unicorn" start-ups in trendy domains such as "voice cloning," a mode of DL associated with deepfakes and fraud (Confino 2024), raise the question of how much malfeasance the public will tolerate in exchange for vague promises of "abundance." The most potentially beneficial data-driven technologies—for example, development of drug treatments or climate efficiencies—are seldom foregrounded in the big tech agenda since the advancement of these projects requires serious investment in human expertise and high-quality data that lies beyond Silicon Valley's doorstep. Such focused technologies, that is to say, are not low-hanging fruit for the corporate giants now vying to turn their access to copious internet babble into new trillion-dollar markets.

Comparing the CALO era to our own day, one observes that, at a time when enormous resources pour into the ongoing generative "craze," the technology industry has yet to produce a robust assistant ("AI"-powered or otherwise) that can securely and reliably handle the cumbersome tasks that dominate modern life (e.g., collecting the email addresses for a group of recipients, scheduling a meeting, providing the information necessary for a doctor's visit, or even renewing a library book). As we have seen, that is so in part because (1) automating such tasks requires painstaking engineering, (2) the digital landscape is dominated by surveillance empires that demand data expropriation as the price of admission, and (3) these monopolistic enterprises have few incentives to design costly tools simply because the billions of people already using their platforms would find them helpful.[49] Thus, instead of purpose-built and secure engineering, the turn to large models has led the most powerful corporations to double down on data accumulation and surveillance. Moreover, the more data-hungry and resource-intensive their products become, the more likely that these companies will impede alternatives and extend their hegemonies far into the future.

These financial incentives reinforce the endorsement of a strong data positivism that endows the industry's commercial endeavors with the gloss of scientific rigor. A decade ago, such positivism made it possible to believe that massive data sets

---

See Ramel 2024 for discussion of a study that documented "disconcerting trends for [the] maintainability" of AI-generated code.

[49] See Doctorow 2023 for an insightful account of how platforms attract users by providing them with high-value services but then change priorities to extract profit in a process that typically leads to "enshittification."

constituted effective totalities that could "solve" language through the methods of data-driven signal processing. With the advent of generative transformers, it made it possible to believe that such statistical techniques could simulate human reasoning and creativity, deliver existing knowledge to human learners, and produce new knowledge or humanlike poiesis. Now, with the chatbotization of generative models, the same mentality aims to persuade the public that human communicative expression simply *is* data-driven signal processing—that writing or art-making is not process but product, not provenance but output.

Thus, while people may still need to paste in email addresses by hand, "superintelligent" machines, they are told, will spare them the burden of high-friction cognitive tasks such as reading books, conducting research, or writing their own correspondence—to say nothing of the time-consuming practices of learning, critical thinking, and dialogue that those tasks have historically supported. From the vantage of data positivism, the process of datafication effectuates an onto-epistemically level playing field.[50] This means that (for data positivists) the act of researching and writing a book—or even the act of carefully reading and commenting on the book's contents—becomes effectively indistinguishable from the act of entering a file for the book or its sources into a chatbot interface for the purpose of prompting some relevant output. Behind such leveling, we contend, is an influential but deceptive ideology of frictionless *knowing* that conflates a person's ability to access (some version of) "past 'lessons'" through a convenient tool, with a person's "full and active 'awareness'" of that modeled content. Needless to say, it is only through direct interaction with both kinds of "experience" that human beings acquire the lived knowledge that helps them to engage and enrich a plural world of other situated people, places, objects, and ideas.[51] Though certain tasks can and arguably should be as frictionless as possible (e.g., filling out a common form), other modes of ostensible "frictionlessness" (e.g., publishing machine-generated "research" as if it were reliable) are clearly misguided and dangerous.

That is not to say that a machine-generated summary may not sometimes be useful (just as a machine-generated translation is sometimes useful). The history of writing, from clay tokens, papyrus, and typewriters to word processors, spell-check, and autocomplete, has always been technologically mediated; and it has always encompassed functional genres of many kinds (including forms, reports,

---

[50] Bode and Goodlad (2023b) define *datafication* as "the process or practices that render a vast multiplicity of objects, locations, activities, and conditions legible *as* data."

[51] It is once again worth clarifying that LLMs are not search engines (providing links to external content) or curated databases. They are, rather, models of undocumented training data: thus, their convenient access to "past lessons" may garble the content of knowledge observed in training data, exclude the most up-to-date and reliable knowledge, or inject false or confabulated information as if it were knowledge. That is true even despite the now frequent use of retrieval augmented generation (RAG), (a technique also discussed in Allison and DeRewal's thinkpiece in part 2 of this special issue), which improves the results of a pre-trained LLM by first querying special databases that contain updated information and then instructing the LLM to respond using the content so obtained. For an example of the major research on this now common technique, see Lewis et al. 2020. On the limitations of RAG-based systems in the legal domain in precluding confabulations (with a reported rate of "hallucinations" between seventeen and thirty-three percent of the time, see Magesh et al. 2024.

advertisements, memoranda, and ersatz digital content produced for search engine optimization).[52] There is no question that many formulaic tasks—including the drafting of repetitive emails—could be partially automated with time-saving benefit to people in many circumstances—just as carefully engineered and purpose-built tools for making a plane reservation, securely gathering email addresses, or booking a doctor's appointment would benefit many people. But at a time when chatbots are being marketed simultaneously as all-purpose knowledge and content systems and AI "companions"— purposes for which they are fundamentally "misaligned," it is important to consider this simple fact: the high-quality human-generated text that LLM developers relish for training their systems did not enter the world through people who took up their tablets, quills, and laptops to predict plausible outputs or resolve ambiguities for practical tasks of data-driven signal processing. That is part of why Ilia Shamailov et al. (2023) believe that training new LLMs on the outputs of existing systems becomes increasingly nonviable as the overproduction of probable content "poisons reality" and leads to "model collapse." What data positivists appear to have forgotten—or perhaps never knew—is that writing is a key dimension of human poiesis: an intersubjective practice rooted in human ontogeny but cultivated through the sustained experience of materially embedded, socially situated, and affectively embodied communicative relations. As bell hooks (1999:13) puts it, "We do not write because we must. . . . We write because language is the way we keep a hold on life."

In her contribution to part 2 of this special issue, a meditation on writing's "lifecycle," Lauren Goodlad argues that those developing, promoting, and even teaching with text generators have often thought very little about what writing is. Although high-quality human writing may involve any number of technologies, commercial purposes, and conditions of possibility, its fundamental signature is the use of written language to share situated reality with other communicative subjects (implied or explicit). According to Goodlad, these criteria stand whether the context in play is a novelist addressing implied readers, a journalist reporting the news, a letter writer reaching out to a friend, a diarist recording the day's impressions, an official issuing public service announcements, a castaway penning a message in a bottle, or—as in Bush's memex—an investigator applying research and critical skills. Thus, Goodlad argues, writing-as-output, even when qualitatively strong, is never the core human purpose of human writing. Rather, the core human purpose (and distinctive signature) of human writing is to constitute writers as writing subjects.

It follows that writing subjects do not directly reproduce registered perceptions of the kind encountered when a person (or puppy) smells food or a mechanical sensor detects the presence of light, or when LLMs cluster vast web corpora into probable word sequences. Rather, through writing, humans invite other people to join them in attending to particular impressions or ideas from situated perspectives that the writing conveys.[53] Think of the speaker of Matthew Arnold's (1867) "Dover Beach," who

[52] For a sample of the diverse and cross-disciplinary body of work on the technological mediation of writing, see, for example, Martin 1994; Warschauer 2010; Balsamo 2011; L. Johnson and Sullivan 2020.

[53] This description of the intersubjective situation for writing builds on the work of Michael Tomasello, especially Tomasello 1999 and Tomasello 2019. For Tomasello, all human communication derives from a

beckons his addressee to join him in sustained reflection on a disenchanting world: "Come to the window" and "Listen!" (lines 6, 9). Such acculturated experiences in (and of) a plural world—the "past 'lessons'" and "full and active 'awareness'" of which Williams writes—are fundamental not only to writing but also to the making of art, the exercise of judgment, and, ultimately, the sharing of reality through any medium.

In the present issue, several contributions resonate with these and other ideas on how human communicative practices compare to machine-generated content. In his essay, data scientist Rafael Alvarado locates the LLMs of today in a structuralist shift away from "the ethnographer's focus on the shared situation" of speakers and toward a "distributional hypothesis" that focuses on "the latent network of words and meanings" within established textual archives. By operationalizing this (implicitly positivistic) hypothesis, LLMs effectively test a distributional approach to meaning that hitherto was largely theoretical. According to Alvarado, LLMs confirm the epistemic affordances of digital distributions when, for example, they execute the formal criteria of poetic genres. But when LLMs visibly stumble over questions of truth-telling, the issue is not simply their well-known tendency to "hallucinate." More noteworthy still is that even when LLMs generate "true" statements, their truths are "unjustified" since the model has no means for validating their claims. Thus, in ways that resonate with Schneider's case for a sociolinguistic approach to communicative practice, Alvarado, in dialogue with Lucy Suchman (1987) and Marshall Sahlins (1985), calls for a return to "the rich ethnographic and ethnohistoric understandings of language."

Literary critic Chloë Kitzinger's thinkpiece comparably notes how text-generating chatbots have prompted a number of high-profile commentators to turn to Plato's *Phaedrus* for insights. Plato's text famously stages a dialogue about the invention of writing as a new technology—one to which Derrida turned in his influential essay "Plato's Pharmacy" (1968), Because they mistakenly regard *Phaedrus* as a "cautionary tale about the futility of distrusting technology," Kitzinger argues, commentators such as the *New York Times*' Ezra Klein mistakenly anoint OpenAI's chatbot as a "successor to the mantle of writing." A better understanding of *Phaedrus* and LLMs, Kitzinger believes, will recognize, as Derrida did, that Plato's text is not a reactionary plaint about human memory loss but rather a case for "active thought and dialogue" as the medium of shared wisdom. Hence, it is not writing that Plato's (writerly) dialogue seeks to avert but the decline of linguistic expression into mimicry. The most insightful Derridean interpretation of LLMs, therefore, will not settle for the truism that these systems produce text without writers; rather, LLMs "produce texts without *writing*." Today's text generators lack the active "free play" with language that, for Derrida, marks the contents of writing's "pharmacy."

It follows that the best digital infrastructures for human writing enable human users by amplifying and concretizing their interactive role in crafting trains of contemplation and rendering this situated experience in shareable form. The point in

uniquely human ability to point, a performative gesture that summons joint attention so as to develop the onto-epistemic and socio-cognitive conditions necessary for shared understanding and collaboration. See also Goodlad's description of the "action" tradition in language theory in her essay in part 2 of this special issue.

making that claim is not to salvage a sovereign monad whose supposed "mastery" over the world affirms the Cartesian worldview or to inflate the powers of a subject-centered autonomy that has never actually existed. To the contrary, writing subjects are, by definition, profoundly embedded and fully immersed in a pluralistic and more-than-human world. Indeed, it is precisely because the textual poiesis that constitutes writers as writing subjects emerges through technologies and technical practices of many kinds, that it matters so much how these tools are designed and with what specific goals in mind.

That is why Kyle Booten's piquant essay in this special issue explores "the design space of interactive systems" that "challenge writers to be stronger and more limber." Locating automated text generators in a long history of proletarianization, Booten, a literary scholar and digital humanist, proposes that we need not retreat to pen and paper in order to ensure tools that "strengthen our minds" and "connect them to other minds." The alternative Booten proposes turns the "AI" myth of frictionless knowing and poiesis on its head: designing "word-gyms" rather than "word-factories." Just as writing was a "good bargain" for human communication in Plato's time (and since), so computational tools that, for example, "goad the writer toward more interesting syntax" can preserve the vigorous play at the heart of writing's ongoing project. Implicit in Booten's seizing the gym as metaphor is a key perception for this introduction and special issue: that *friction*—the experience of difficulty, resistance, and the passage of time—in writing, as in other modes of human poiesis, is an inescapable index of the condition of being conscious, embodied, interactive, and alive.[54]

In contrast to a culture that recognizes friction as integral to the temporality and lived materiality of human cultural poiesis, imagine a near future in which Chatbot-K becomes the dominant multimodal interface for efficient communications: inculcated into populations in or before primary school, sustained through higher education, enforced at the workplace, and marketed on various platforms for research, leisure, politics, companionship, and love. Centralized, profitable, and unaccountable, such an all-purpose bot—devoted to promoting human mimicry and the myth of frictionless knowing—could be troubling in several ways. A Chatbot-K on this scale could amplify biases, misinformation, and political dysfunction, and help malicious users amp up demeaning content (with especially damaging impact on the vulnerable and marginalized). As Wong writes in her response to Sam Bowman in part 2 of this special issue, such a technology could pose a threat to human rights simply because "generative systems—and AI and datafication more generally—push the boundaries of what it means to be human" so significantly that "human rights" become "immediately relevant as the moral architecture through which to understand these changes." For example, endowed with multimodal features trained on the work of authors and artists past, present, and future, a Chatbot-K could enervate human creativity as it foments homogeneity while undermining the conditions for diverse human craft.

[54] For strong meditations on frictionlessness as a harmful myth, see also Benjamin 2022: 162–63; Elam 2023; Goodlad and Baker 2023.

Moreover, a normatively "reinforced" Chatbot-K might foster a whitewashed establishment rhetoric of neoliberal precepts, political quiescence, and corporate groupthink. In such a monoculture "AI Tubman" might stand in for the history of civil rights; technical debt could take the form of diminished "active 'awareness'"; and data positivism could purport to justify an increasingly unequal society that degrades human expertise (even while ensuring high-quality professional services for those with the resources to insist on nothing less for themselves and their families). A Chatbot-K might market itself as an AI-driven *Wunderwaffe* or all-purpose machinery for "solutions" of every kind: factoids for the learner, bromides for the sick, friendship for the lonely, research hacks for the overwhelmed, and tips for the ambitious and tenacious.[55] Acclaimed as the product of scientific genius—sprung from the heads of AI godfathers!—such "AI" would in reality rely on never-ending surveillance; probabilistic scoring; the consumption of vast resources; and a crowdworker underclass numbering into the billions.

Although we coauthors do not by any means predict this outcome, we know that the hard work of resisting it—strengthening democracy; caring for the earth; and nurturing its myriad cultures, objects, and ways of living (inside and outside of critical technologies)—could be the work of a generation or more. It is perhaps above all the work of our students—some of whom we hope are reading these pages. Like Estrin, we believe that piercing the "illusion of market forces" is crucial for this enterprise. But we also concur with digital humanist Megan Ward: "An effective critique of AI," she writes in her review essay in the current issue, "must do more than bring . . . disavowed values and knowledge practices to light." It must also delineate "alternate futures." As admiring readers of Ruha Benajmin's (2022: 54) case for viral justice, we also know that "grassroots efforts" and local alliances are the springs of such critical worldbuilding.

Clearly one pathway to alternate futures is to regulate "authoritarian intelligence" and rein in its underlying political economy—in much the way that child labor, railroad monopolies, food, automobiles, aviation, and television were regulated in the past. In the United States, in the last year alone, diverse initiatives and policymaking bodies (including the AI Now Institute, the Algorithmic Justice League, and the DAIR Institute) made progress in educating the public, the media, and the US Congress. Nascent

[55] For an important analysis of the "epistemic risks" of deploying chatbots for a range of research tasks—not only as tools for *writing* research but also as automated analysts, hypothesizers, and substitutes for human research subjects—see Messeri and Crockett 2024. The authors describe four main research roles through which AI-enthusiasts uphold the alleged research capacities of LLMs ("Oracles," "Surrogates," "Quants," and "Arbiters"). As confidence in each of these roles becomes mutually reinforcing, researchers may mistake a narrow monoculture for a superior set of methods which they falsely believe to be more objective and universal than alternative perspectives—a situation in which more *productivity* results in less learning, knowledge, or understanding (see also Goodlad's essay in part 2 of this special issue). The results could deteriorate the efforts to diversify research which have produced incremental improvements over decades. For the compatible argument that LLMs are incapable of portraying demographically diverse research subjects without significant distortion and flattening, see Wang, Morgenstern, and Dickerson 2024. For the compatible argument that the knowledge-building of AI ethics is entrenched in epistemic power imbalances that "delegitimize and marginalize embodied and lived experiences," see David Gray Widder 2024. See also Newfield 2023 on epistemic hierarchies between science and technology disciplines on the one hand, and humanities and social sciences on the other.

programs for teaching critical AI literacy to students, educators, and communities are offering robust alternatives to industry-dominated discourse, doomerism, and hype. In her review of *Data Conscience* (2023) by the computer scientist Brandeis Marshall, digital humanist and data scientist Catherine D'Ignazio praises the book's ambitious proposal to reorient the sphere of law and regulation toward active prevention of harms. That forward-moving outlook flickers in Emily M. Bender's perception (in a dialogue with Ted Chiang in this special issue) that lawmakers are increasingly extending their conversations beyond "talking to corporations."[56] **{Au: This sentence does not appear in the interview -CORRECTED }** Moreover, as Chiang reminds us in the same conversation, it will take more than diminishing the market for paid writing and art to extinguish the inclination toward human poiesis. "The reality of being a writer," Chiang attests, "is that you don't know that you'll ever connect with someone, and you do it anyway."

As legal scholar Ellen P. Goodman explains in her thinkpiece in this special issue, regulatory strategies are complicated by the fact of AI's myriad forms. Common analogies to the oversight of nuclear weapons, pharmaceuticals, and/or environmental pollution provide useful templates—but none is a wholly compatible or comprehensive strategy for "AI." Thus, in the months and years ahead, writes Goodman—who is now senior adviser for algorithmic justice to the National Telecom and Information Administration—AI regulators will need to look deeply into how these systems function and who bears (or should bear) the costs of their harms. We hope that some of the people researching those questions are also reading these pages. Insofar as their work calls for new conversations, protocols, and regulatory levers suitable for accountability across diverse domains, we hope the essays, thinkpieces, and reviews in this two-part special issue will help to season that work.

One of the most remarkable recent victories for collaborative action was the 2023 Writers Guild of America/SAG-AFTRA strike—a historic struggle against the use of technology to undermine solidarity and economic justice.[57] As we write, elite technology workers, still reeling from the postpandemic decimation of their ranks in 2023, have been told that the competition to develop AI will require more of the same (see, e.g., Heath 2024). To be clear: these jobs have not been significantly automated. Nonetheless, for investors to continue bankrolling commercial "AI" technologies that

[56] On a range of regulatory initiatives already in motion see, for example, *Financial Times* 2024; Sorkin et al. 2024. For details of the Justice Department's antitrust suit against Google, see the Biden Administration's executive order "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Exec. Order No. 14110, 88 Fed. Reg. 75191 (Nov. 1, 2023). In July 2023, the Federal Trade Commission "opened an expansive investigation into OpenAI," citing concerns over "personal reputations" and data privacy (Zakrzewski 2023). Cat Zakrzewski's article also cited Senate Majority Leader Charles Schumer predicting that "new AI legislation" would be coming in a matter of months. As the *New York Times* put it, the FTC was moving exceptionally fast (Kang and Metz 2023). For an example of vigorous public interest advocacy at work, see West 2023, on the increasing use of AI to manipulate pricing.

[57] As tech columnist Merchant (2023c) explained, the strike prevented studios from using "AI" "as leverage against writers, both as a threat and as a means to justify offering lowered rewrite fees."

lose millions of dollars as a matter of course, the highly skilled people at work on these products are pressed to narrow their focus and do more with less.[58]

Can we—by which we mean the readers of this special issue—help to actuate worldbuilding in common with colleagues in hard-hit fields including creative workers, journalists, editors, students, office staff, gig workers, and crowdworkers, among them? Can we pit the well-lubricated fantasy of frictionless knowing against the technical practices and "awareness" necessary for limber and connected communities?

We hope the answer is yes.

According to Ward, the move beyond critique and toward alternative futures "effectively requires a transdisciplinary approach"—one that is admittedly difficult to implement given structural barriers and "the yawning gap in perceived value between STEM and non-STEM fields." Hence, even as she recognizes what Christopher Newfield (2023) described as the core inequalities between STEM and non-STEM domains in *Critical AI*'s first issue, Ward argues that encounters across these uneven boundaries must be strengthened. We agree. As Bode and Goodlad (2023b) put it, critical AI studies is less a defined methodology than a community of multifaceted interdisciplinary practice premised partly on the belief that "spaces of dissensus" can activate the future. It is also a practice of teaching critical AI literacies and—in doing so—exploring the insights of design justice principles (Design Justice Network 2018).

As those principles show, a world in which technology serves the public interest is one that invites diverse people to share knowledge, experience, and technological needs and practices. It is a world that embraces friction as the frisson, and sometimes the heartbreak, that comes with "full and active 'awareness'"—the state of being conscious, embodied, interactive, and alive.

Lauren M. E. Goodlad is the editor of *Critical AI* and the chair of the Critical AI @ Rutgers initiative.

Matthew Stone is professor in the Department of Computer Science and Center for Cognitive Science at Rutgers. He has served as program chair for the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) and general chair of the meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). He was chair of the computer science department from 2019 to 2023 and

[58] Though hardly comparable to the taxi drivers whose livelihoods were severely damaged by technological "disruption"—or even to their lesser counterparts in contract work (see, e.g., Sheng et al. 2019)—prosperous technology workers are among the most privileged to experience a wave of insecurity that is perceived to coincide "with the rise of potentially job-killing" AI (Nagpaul 2024). At least for now, automation is less a direct factor in layoffs at high-profile technology companies, as is the strategic focus on "generative" products and (as above) the appeasement of investors. As sociologist John David Skrentny (2024) argues in a recent opinion essay, the career prospects for STEM majors are as broken as those in other professions because "investors repeatedly reward employers for treating STEM grads like fast fashion—discarded when . . . no longer appealing," or "for deploying STEM skills in lucrative yet harmful business models." See also Rotman 2023 on the AI economy: the record of the last decade's advances "in improving prosperity and spurring widespread economic growth is discouraging. Although a few investors and entrepreneurs have become very rich, most people haven't benefited."

has helped to organize interdisciplinary training programs at Rutgers in Perceptual Science, Data Science, and Socially Cognizant Robotics. A major theme of his research is embodied communication, and the potential for gesture, facial expressions, demonstrations, diagrams, and other coverbal actions to enhance human-computer communication. His work was recognized with the IFAAMAS Influential Paper Award in 2017. He serves on the *Critical AI* editorial collective.

Acknowledgments: We are grateful to Mark Sammons, Katie Conrad, Chloë Kitzinger, and Christopher Newfield for helpful feedback on the introduction and to Sabrina Burns, Andi Craciun, Kelsey Keyes, and Jai Yadav for technical assistance with the works cited. We acknowledge the support of NEH RZ-292740–23 (Design Justice Labs) in the preparation of this essay and are grateful as well to Mellon-CHCI for the Design Justice AI global humanities institute funding that has expanded our research into the impact of LLMs on local languages. Stone's work on this essay was also supported by NSF IIS-211926 and DGE-202162 and by a sabbatical leave award from Rutgers. Lauren Goodlad's work on this essay was also supported by the Dean of Humanities office at Rutgers.

Works Cited

Abebe, Rediet, and Maximilian Kasy. 2021. "The Means of Prediction." *Boston Review*, May 20. https://www.bostonreview.net/forum_response/the-means-of-prediction/.

Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. "Persistent Anti-Muslim Bias in Large Language Models." Preprint, last revised January 18. https://doi.org/10.48550/arXiv.2101.05783.

Alexa Prize Team. 2023. "Alexa Prize SocialBot Grand Challenge 5 Winners Announced." *Amazon Science*, September 12. https://www.amazon.science/alexa-prize/socialbot-grand-challenge/2022.

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

Angwin, Julia. 2023. "The Internet Is About to Get Much Worse." *New York Times*, September 23. https://www.nytimes.com/2023/09/23/opinion/ai-internet-lawsuit.html.

Anthony, Andrew. 'Eugenics on Steroids': the Toxic and Contested Legacy of Oxford's Future of Humanity Institute." *Guardian,* April 28. https://www.theguardian.com/technology/2024/apr/28/nick-bostrom-controversial-future-of-humanity-institute-closure-longtermism-affective-altruism.

Appel, Gil, Juliana Neelbauer, David A. Schweidel. 2023. "Generative AI Has an Intellectual Property Problem." *Harvard Business Review*, April 7. https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.

Arnold, Matthew. 1867. "Dover Beach." Poetry Foundation. https://www.poetryfoundation.org/poems/43588/dover-beach.

Associated Press. 2023. “Michael Cohen Says He Unwittingly Sent AI-Generated Fake Legal Cases to His Attorney.” NPR, December 30. https://www.npr.org/2023/12/30/1222273745/michael-cohen-ai-fake-legal-cases.

Bai, Xuechunzi, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. “Measuring Implicit Bias in Explicitly Unbiased Large Language Models.” *arXiv preprint arXiv:2402.04105*.

Bale, Ajay Sudhir, R. B. Dhumale, Nimisha Beri, Melanie Lourens, Raj A. Varma, Vinod Kumar, Sanjay Sanamdikar, and Mamta B. Savadatti. 2024. “The Impact of Generative Content on Individuals Privacy and Ethical Concerns.” https://ijisae.org/index.php/IJISAE/article/view/3503.

Balsamo, Anne. 2011. *Designing Culture: The Technological Imagination at Work*. Durham, NC: Duke University Press. https://doi.org/10.1215/9780822392149.

Bender, Emily M. 2022. “On *NYT Magazine* on AI: Resist the Urge to Be Impressed.” *Medium*, May 2. https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd.

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. https://doi.org/10.1145/3442188.3445922.

Bender, Emily M., and Alexander Koller. 2020. “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data.” In *Proceedings of the Fifty-Eighth Annual Meeting of the Association for Computational Linguistics*, 5185–98. Stroudsburg, PA: Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.463.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. New York: Polity.

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

University Press.

Bhasker, Shashank, Damien Bruce, Jessica Lamb, and George Stein. 2023. “Tackling Healthcare’s Biggest Burdens with Generative AI.” McKinsey & Company, July 10. https://www.mckinsey.com/industries/healthcare/our-insights/tackling-healthcares-biggest-burdens-with-generative-ai.

Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2022. “Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale.” Preprint, submitted November 7. https://doi.org/10.48550/arXiv.2211.03759.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. “Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes.” Preprint, submitted October 5. https://doi.org/10.48550/arXiv.2110.01963.

Birhane, Abeba, and Inioluwa Deborah Raji. 2022. “ChatGPT, Galactica, and the Progress Trap.” *Wired*, December 9. https://www.wired.com/story/large-language-models-critique/.

Bode, Katherine, and Lauren M. E. Goodlad, eds. 2023a. “Data Worlds.” Special issue, *Critical AI* 1, nos. 1–2. https://read.dukeupress.edu/critical-ai/issue/1/1-2.

Bode, Katherine, and Lauren M. E. Goodlad. 2023b. “Data Worlds: An Introduction.” In Bode and Goodlad 2023a. https://doi.org/10.1215/2834703x-10734026.

Bommasani, Rishi, et al. 2021. “On the Opportunities and Risks of Foundation Models.” Preprint, last revised August 18. https://doi.org/10.48550/arXiv.2108.07258.

Broussard, Meredith. 2023. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. Cambridge, MA: MIT Press.

Bubeck, Sébastien, et al. 2023. “Sparks of Artificial General Intelligence: Early Experiments With GPT-4.” Preprint, submitted March 22. 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* 81:77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

Byman, Daniel L., Chongyang Gao, Chris Meserole, and V. S. Subrahmanian. 2023. *Deepfakes and International Conflict: A Research Report*. Washington, DC: Brookings Institution. https://www.scholars.northwestern.edu/en/publications/deepfakes-and-international-conflict-a-research-report.

Cai, Kenrick. 2023. “AI Video Startup HeyGen Launches Near-Instant Avatar Generator, Adds $5.6 Million in Funding.” *Forbes*, November 29. https://www.forbes.com/sites/kenrickcai/2023/11/29/ai-video-startup-heygen-launches-near-instant-avatar-generator-adds-56-million-in-funding/?sh=7de1cce86782.

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

Chen, Claire. 2023. “AI Will Transform Teaching and Learning. Let’s Get It Right.” *Stanford HAI*, March 9. https://hai.stanford.edu/news/ai-will-transform-teaching-and-learning-lets-get-it-right.

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. “ChatGPT Is a Blurry JPEG of the Web.” *New Yorker*, February 9. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Chollet, François (@fchollet). 2023. “The first panic over imminent AGI was circa 2013 about Atari Q-learning by DeepMind.” X, November 23, 10:18 p.m. https://twitter.com/fchollet/status/1727798735676350553.

Christian, Jon. 2023. “Amazing ‘Jailbreak’ Bypasses ChatGPT’s Ethics Safeguards.” *Futurism*, February 4. https://futurism.com/amazing-jailbreak-chatgpt.

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

CIO Coverage. n.d. “OpenAI’s ChatGPT Reportedly Costs $100,000 a Day to Run.” https://www.ciocoverage.com/openais-chatgpt-reportedly-costs-100000-a-day-to-run/. **{Au: Please provide a last-updated date or access date.}**

Climate Action against Disinformation Coalition. March 2024. *Artificial Intelligence Threats to Climate*. https://foe.org/wpcontent/uploads/2024/03/AI_Climate_Disinfo_v6_031224.pdf.

Cole, Samantha. 2023. “Largest Dataset Powering AI Images Removed after Discovery of Child Sexual Abuse Material.” *404 Media*, December 20. https://www.404media.co/laion-datasets-removed-stanford-csam-child-abuse/.

Confino, Paolo. 2024. “Another AI Unicorn? $80 Million Series B Led by Andreessen Horowitz Yields a $1.1 Billion Valuation, Source Says.” *Fortune*, January 22. https://fortune.com/2024/01/22/ai-unicorn-andreesen-horowitz-eleven-labs-venture-capital-series-b-tech/.

Costanza-Chock, Sasha. 2020. *Design Justice: Community-Led Practices to Build the Worlds that We Need.* Cambridge, MA: MIT Press.

Coulter, Martin. 2023. “AI Experts Disown Musk-Backed Campaign Citing Their Research.” *Reuters*, April 5. https://www.reuters.com/technology/ai-experts-disown-musk-backed-campaign-citing-their-research-2023-03-31/.

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Davis, Wes. 2023. “Sarah Silverman Is Suing OpenAI and Meta for Copyright Infringement.” *Verge*, July 9. https://www.theverge.com/2023/7/9/23788741/sarah-silverman-openai-meta-chatgpt-llama-copyright-infringement-chatbots-artificial-intelligence-ai.

De Cremer, David, Nicola Morini Bianzino, and Ben Falk. 2023. “How Generative AI Could Disrupt Creative Work.” *Harvard Business Review*, April 13. https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work.

De Vynck, Gerritt, 2024. “OpenAI’s News Deals Continue, with Vox and the Atlantic Signing On.” *Washington Post*, May 29.

https://www.washingtonpost.com/technology/2024/05/29/openai-vox-the-atlantic-news-deals/.

Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. "On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet." *Big Data and Society* 8, no. 2: 205395172110359. https://doi.org/10.1177/20539517211035955.

Derrida, Jacques. 1968. "Plato's Pharmacy." In *Dissemination*, translated by Barbara Johnson, 61–171. Chicago: University of Chicago Press.

Design Justice Network. 2018. "Design Justice Network Principles." https://designjustice.org/read-the-principles. **{Au: Please provide a last-updated date, if available, or access date (with month and day)}**

Devinney, Hannah, Jenny Björklund, and Henrik Björklund. 2022. "Theories of 'Gender' in NLP Bias Research." In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2083–102. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3534627.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge, MA: MIT Press.

Doctorow, Cory. 2020. "How to Destroy Surveillance Capitalism." *Medium*, August 26. https://onezero.medium.com/how-to-destroy-surveillance-capitalism-8135e6744d59.

Doctorow, Cory. 2023. "Tiktok's Enshittification." *Pluralistic: Daily Links from Cory Doctorow*, January 21, 2023. https://pluralistic.net/2023/01/21/potemkin-ai/#hey-guys.

Doctorow, Cory. 2024. "'Enshittification' Is Coming for Absolutely Everything." *Financial Times*, February 8. https://www.ft.com/content/6fb1602d-a08b-4a8c-bac0-047b7d64aba5.

Drahl, Carmen. 2023. "AI Was Asked to Create Images of Black African Docs Treating White Kids. How'd It Go?" *NPR*, October 6. https://www.npr.org/sections/goatsandsoda/2023/10/06/1201840678/ai-was-asked-to-create-images-of-black-african-docs-treating-white-kids-howd-it-.

Dzieza, Josh. 2023. "Inside the AI Factory: The Humans That Make Tech Seem Human." *Intelligencer*, June 20. https://nymag.com/intelligencer/article/ai-artificial-intelligence-humans-technology-business-factory.html.

Dziri, Nouha, et al. 2023. "Faith and Fate: Limits of Transformers on Compositionality." Preprint, submitted May 29. https://doi.org/10.48550/arxiv.2305.18654.

Edelman, David C., and Mark Abraham. 2023. "Generative AI Will Change Your Business: Here's How to Adapt." *Harvard Business Review*, April 12. https://hbr.org/2023/04/generative-ai-will-change-your-business-heres-how-to-adapt.

Edwards, Benj. 2023a. “Google’s Best Gemini AI Demo Video Was Fabricated.” *Ars Technica*, December 8. https://arstechnica.com/information-technology/2023/12/google-admits-it-fudged-a-gemini-ai-demo-video-which-critics-say-misled-viewers/.

Edwards, Benj. 2023b. “OpenAI Introduces GPT-4 Turbo: Larger Memory, Lower Cost, New Knowledge.” *Ars Technica*, November 6. https://arstechnica.com/information-technology/2023/11/openai-introduces-gpt-4-turbo-larger-memory-lower-cost-new-knowledge/.

Elam, Michele. 2023. “Poetry Will Not Optimize; or, What Is Literature to AI?” *American Literature* 95, no. 2: 281–303. https://doi.org/10.1215/00029831-10575077.

EPIC. n.d. “Enforcement of Privacy Laws.” Electronic Privacy Information Center. https://epic.org/issues/data-protection/enforcement-of-privacy-laws/. **{Au: Please provide a last-updated date or access date.}**

Estrin, Judy. 2023. “The Case against AI Everything, Everywhere, All at Once.” *Time*, August 11. https://time.com/6302761/ai-risks-autonomy/.

Farhat, Eamon. 2024. “Cryptocurrency, AI Electricity Demand Seen Doubling in Three Years.” *Bloomberg*, January 24. https://www.bloomberg.com/news/articles/2024-01-24/cryptocurrency-ai-electricity-demand-seen-doubling-in-three-years.

Financial Times. 2024. “Global Regulators’ Assault Is Already Crimping Big Tech’s Prospects.” January 23. https://www.ft.com/content/0a753971-377e-4ada-951f-f23a620106e4.

Fletcher, Harry. 2023. “Bill Gates Says That Technology Can Help Make a Three Day Work Week Possible.” *Indy100*, November 23. https://www.indy100.com/science-tech/bill-gates-three-day-week.

Franklin, Seb. 2023. “Data/Dispossession.” *American Literature* 95, no. 2: 321–35. https://doi.org/10.1215/00029831-10575105.

Funk, Allie, Adrian Shahbaz, and Kian Vesteinsson. 2023. “The Repressive Power of Artificial Intelligence.” Freedom House. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence. **{Au: Please provide a last-updated date, if available}**

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

Gardizy, Anissa, and Aaron Holmes. 2024. “Amazon, Google Quietly Tamp Down Generative AI Expectations.” *Information*, March 12. https://www.theinformation.com/articles/generative-ai-providers-quietly-tamp-down-expectations.

Gardizy, Anissa and Wayne Ma. 2023. “Microsoft Readies AI Chip as Machine Learning Costs Surge.” *Information*, April 18.

https://www.theinformation.com/articles/microsoft-readies-ai-chip-as-machine-learning-costs-surge

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." DAIR 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." Preprint, last revised September 25. https://doi.org/10.48550/arXiv.2009.11462.

Giansiracusa, Noah. 2023. "Noah Giansiracusa on 'You Can Have the Blue Pill or the Red Pill, and We're Out of Blue Pills' in The New York Times (3/24/2023)." *Critical AI Wall of Shame*, April 3. https://criticalai.org/2023/04/06/noah-giansiracusa-on-you-can-have-the-blue-pill-or-the-red-pill-and-were-out-of-blue-pills-in-the-new-york-times-3-24-2023/.

Giridharadas, Anand. 2018. *Winners Take All: The Elite Charade of Changing the World*. New York: Knopf.

Gitelman, Lisa, ed. 2013. "'Raw Data' Is an Oxymoron." Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001.

Goldenberg, Suzanne. 2005. "Why Women Are Poor at Science, by Harvard President." *Guardian*, January 18. https://www.theguardian.com/science/2005/jan/18/educationsgendergap.genderissues.

Goldman, Sharon. 2023. "AI Pioneers Hinton, Ng, LeCun, Bengio Amp Up X-Risk Debate." *VentureBeat*, October 31. https://venturebeat.com/ai/ai-pioneers-hinton-ng-lecun-bengio-amp-up-x-risk-debate/.

Goode, Lauren. 2024. "Google Prepares for a Future where Search Isn't King." *Wired*, February 9. https://wired.me/business/google-for-future-where-search-isnt-king/.

Goodlad, Lauren M. E. 2023. "Editor's Introduction: Humanities in the Loop." *Critical AI* 1, nos. 1–2. https://doi.org/10.1215/2834703x-10734016.

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/.

Goodlad, Lauren M. E., and Kathryn Conrad. 2024. "Teaching Critical AI Literacies." NORRAG Policy Insights #4 *AI and Digital Inequities*. March, https://resources.norrag.org/resource/845/policy-insights-ai-and-digital-inequities

Grant, Nico. "Google's A.I. Search Errors Cause a Furor Online." 2024. *New York Times*. May 24. https://www.nytimes.com/2024/05/24/technology/google-ai-overview-search.html?searchResultPosition=1.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. New York: Harper Business.

Grynbaum, Michael M., and Ryan Mac. 2023. “The Times Sues OpenAI and Microsoft over A.I. Use of Copyrighted Work.” *New York Times*, December 27. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. “The Unreasonable Effectiveness of Data.” *IEEE Intelligent Systems* 24, no. 2: 8–12. https://doi.org/10.1109/MIS.2009.36.

Hanna, Alex, and Emily M. Bender. 2023. “AI Causes Real Harm. Let’s Focus on That over the End-of-Humanity Hype.” *Scientific American*, August 12. https://www.scientificamerican.com/article/we-need-to-focus-on-ais-real-harms-not-imaginary-existential-risks/.

Hao, Karen. 2019. “Training a Single AI Model Can Emit as Much Carbon as Five Cars in Their Lifetimes.” *MIT Technology Review*, June 6. https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/.

Hao, Karen. 2020. “We Read the Paper That Forced Timnit Gebru Out of Google. Here’s What It Says.” *MIT Technology Review*, December 4. https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/.

Hao, Karen. 2023. “Why Won’t OpenAI Say What the Q* Algorithm Is?” *Atlantic*, November 28. https://www.theatlantic.com/technology/archive/2023/11/openai-sam-altman-q-algorithm-breakthrough-project/676163/.

Hao, Karen. 2024. “AI Is Taking Water from the Desert.” *Atlantic*, March 4. https://www.theatlantic.com/technology/archive/2024/03/ai-water-climate-microsoft/677602/.

Hao, Karen, and Charlie Warzel. 2023. “How ChatGPT Fractured OpenAI.” *Atlantic*, November 19. https://www.theatlantic.com/technology/archive/2023/11/sam-altman-open-ai-chatgpt-chaos/676050/.

Harari, Yuval, Tristan Harris, and Aza Raskin. 2023. “You Can Have the Blue Pill or the Red Pill, and We’re Out of Blue Pills.” *New York Times*, March 24. https://www.nytimes.com/2023/03/24/opinion/yuval-harari-ai-chatgpt.html.

Harnad, Stevan. 1990. “The Symbol Grounding Problem.” *Physica D: Nonlinear Phenomena* 42, nos. 1–3: 335–46. https://doi.org/10.1016/0167-2789(90)90087-6.

Hauser, Jeff. 2023. “Larry Summers’ Ascent to OpenAI Board Is Awful News for Humanity.” Revolving Door Project, November 22. https://therevolvingdoorproject.org/larry-summers-openai/.

Hays, Kali. 2023. “Firms like Meta and a16z Admit Having to Pay Billions for Training Data Would Ruin Their Generative AI Plans as They Fight New Copyright Rules.” *Business Insider Nederland*, November 3. https://www.businessinsider.nl/generative-ai-copyright-meta-google-openai-a16z-microsoft.

Heath, Alex. 2024. “Meta’s New Goal Is to Build Artificial General Intelligence.” *Verge*, January 18. https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview.

Heath, Alex, and Elizabeth Lopatto. 2024. “OpenAI Says Elon Musk Wanted ‘Absolute Control’ of the Company.” *Verge*, March 6. https://www.theverge.com/2024/3/5/24091773/openai-response-elon-musk-breach-of-contract-lawsuit.

Heaven, Will Douglas. 2024. “AI for Everything: Ten Breakthrough Technologies 2024.” *MIT Technology Review*, January 8. https://www.technologyreview.com/2024/01/08/1085096/artificial-intelligence-generative-ai-chatgpt-open-ai-breakthrough-technologies/.

Heikkilä, Melissa. 2022. “Dutch Scandal Serves as a Warning for Europe over Risks of Using Algorithms.” *Politico*, April 13. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

Herrman, John. 2024. “ChatGPT and Google Gemini Are Both Doomed.” *Intelligencer*, March 1. https://nymag.com/intelligencer/article/chatgpt-and-google-gemini-are-both-doomed.html.

Hinton, Geoffrey. “‘Godfather of AI’ Warns That AI May Figure Out How to Kill People.” Interview by Jake Tapper. CNN, May 3. YouTube video, 4:10. https://www.youtube.com/watch?v=FAbsoxQtUwM.

Hofmann, Valentin, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. “Dialect Prejudice Predicts AI Decisions about People’s Character, Employability, and Criminality.” Preprint, submitted March 1. https://arxiv.org/abs/2403.00742.

Hofstadter, Douglas R. 1995. “Preface 4 The Ineradicable Eliza Effect and its Dangers.” *Fluid*

Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of

Thought. Basic Books. 155-168.

hooks, bell. 1999. *Remembered Rapture: The Writer at Work*. New York: Holt.

Hsu, Tiffany, and Stuart A. Thompson. 2023. “Disinformation Researchers Raise Alarms about A.I. Chatbots.” *New York Times*, February 8. https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.

Hundt, Andrew, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. 2022. “Robots Enact Malignant Stereotypes.” In *FAccT ’22: Proceedings of the*

2022 ACM Conference on Fairness, Accountability, and Transparency, 743–56. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533138.

Inflection. 2024. "Inflection-2.5: Meet the World's Best Personal AI." March 7. https://inflection.ai/inflection-2-5.

IPCC (Intergovernmental Panel on Climate Change). 2023. "AR6 Synthesis Report: Summary for Policymakers Headline Statements." **{Au: Please provide a last-updated date}** https://www.ipcc.ch/report/ar6/syr/resources/spm-headline-statements/.

Irani, Lilly. 2013. "The Cultural Work of Microwork." *New Media and Society* 17, no. 5: 720–39. https://doi.org/10.1177/1461444813511926.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55, no. 12: 248. https://doi.org/10.1145/3571730.

Jiang, Haimin, Lauren T. Brown, Junfen Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. "AI Art and Its Impact on Artists." In *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–74. New York: Association for Computing Machinery. https://doi.org/10.1145/3600211.3604681.

Johnson, Latrise P., and Hannah Sullivan. 2020. "Revealing the Human and the Writer: The Promise of a Humanizing Writing Pedagogy for Black Students." *Research in the Teaching of English* 54, no. 4: 418–38. https://doi.org/10.58680/rte202030740.

Johnson, Steven, and Nikita Iziev. 2022. "A.I. Is Mastering Language. Should We Trust What It Says?" *New York Times*, April 15. https://www.nytimes.com/2022/04/15/magazine/ai-language.html.

Kabir, Samia, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. "Is Stack Overflow Obsolete? Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions." August 4. https://arxiv.org/abs/2308.02312.

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

Kak, Amba, Sarah Myers West, and Meredith Whittaker. 2023. "Make No Mistake—AI Is Owned by Big Tech." *MIT Technology Review*, December 5. https://www.technologyreview.com/2023/12/05/1084393/make-no-mistake-ai-is-owned-by-big-tech/.

Kalluri, Pratyusha Ria, William Agnew, Myra Cheng, Kentrell Owens, Luca Soldaini, and Abeba Birhane. 2023. "The Surveillance AI Pipeline." Preprint, last revised October 17. https://doi.org/10.48550/arXiv.2309.15084.

Kang, Cecilia, and Cade Metz. 2023. “F.T.C. Opens Investigation into ChatGPT Maker over Technology’s Potential Harms.” *New York Times*, July 13. https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html.

Keane, Webb, and Scott J. Shapiro. 2023. “Deus Ex Machina: The Dangers of AI Godbots.” *Spectator*, July 29. https://www.spectator.co.uk/article/deus-ex-machina-the-dangers-of-ai-godbots/.

Khatri, Chandra, et al. 2018. “Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize.” Preprint, submitted December 27. https://doi.org/10.48550/arXiv.1812.10757.

Khatun, Aisha, and Daniel G. Brown. 2023. “Reliability Check: An Analysis of GPT-3’s Response to Sensitive Topics and Prompt Wording.” Preprint, submitted June 9. https://doi.org/10.48550/arXiv.2306.06199.

Klein, Naomi. 2023. “AI Machines Aren’t ‘Hallucinating.’ But Their Makers Are.” *Guardian*, May 8. https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein.

Knibbs, Kate. 2024. “Scammy AI-Generated Books Are Flooding Amazon.” *Wired*, January 10. https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/.

Knight, Will. 2023a. “OpenAI’s CEO Says the Age of Giant AI Models Is Already Over.” *Wired*, April 17. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

Knight, Will. 2023b. “Some Glimpse AGI in ChatGPT. Others Call It a Mirage.” *Wired*, April 18. https://www.wired.com/story/chatgpt-agi-intelligence/.

Kolodny, Lora. 2023. “Cruise Confirms Robotaxis Rely on Human Assistance Every Four to Five Miles.” *CNBC*, November 6. https://www.cnbc.com/2023/11/06/cruise-confirms-robotaxis-rely-on-human-assistance-every-4-to-5-miles.html.

Konrad, Alex. 2024. *Forbes* (Australia). “AI Unicorn Inflection Abandons its ChatGPT Challenger as CEO Mustafa Suleyman Joins Microsoft.” March 19, https://www.forbes.com.au/news/innovation/ai-unicorn-inflection-abandons-chatgpt-challenger-as-ceo-joins-microsoft/.

Kotek, Hadas, Rikker Dockum, and David Sun. 2023. “Gender Bias and Stereotypes in Large Language Models.” In *CI ’23: Proceedings of The ACM Collective Intelligence* Conference, 12–24. New York: Association for Computing Machinery. https://doi.org/10.1145/3582269.3615599.

Langley, Hugh. 2023. “Google’s Water Use Is Soaring. AI Is Only Going to Make It Worse.” *Business Insider*, July 24. https://www.businessinsider.com/google-water-use-soaring-ai-make-it-worse-data-centers-2023-7.

Latour, Bruno. 1999. “On Recalling Ant.” In “Actor Network Theory and After.” Supplement, *Sociological Review* 47, no. 1: 15–25. https://doi.org/10.1111/j.1467-954x.1999.tb03480.x.

LeCun, Yann (@ylecun). 2023. “Current LLMs are trained on text data that would take 20,000 years for a human to read.” *X*, November 23, 5:33 p.m. https://twitter.com/ylecun/status/1727727093671145978.

Leffer, Lauren. 2023. “The AI Boom Could Use a Shocking Amount of Electricity.” *Scientific American*, October 12. https://www.scientificamerican.com/article/the-ai-boom-could-use-a-shocking-amount-of-electricity/.

Levy, Sharon, Michael Saxon, and William Yang Wang. 2021. “Investigating Memorization of Conspiracy Theories in Text Generation.” In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4718–29. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.416.

Levy, Steven. 2023. “Microsoft’s Satya Nadella Is Betting Everything on AI.” *Wired*, June 13. https://www.wired.com/story/microsofts-satya-nadella-is-betting-everything-on-ai/.

Lewis, Patrick et. al. 2020. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.”

Li, Pengfei, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2023. “Making AI Less ‘Thirsty’: Uncovering and Addressing the Secret Water Footprint of AI Models.” Preprint, submitted April 6. https://doi.org/10.48550/arXiv.2304.03271.

Liedtke, Michael. 2023. “Inflection.ai CEO Mustafa Suleyman Explains How to Catch a Ride on the ‘Coming Wave’ of Technology.” *AP News*, September 11. https://apnews.com/article/artificial-intelligence-inflection-mustafa-suleyman-coming-wave-1f9e66f8ed99a769159d384d19a820e0.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. “TruthfulQA: Measuring How Models Mimic Human Falsehoods.” Preprint, submitted September 8. https://doi.org/10.48550/arXiv.2109.07958.

Lu, Kaiji, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2018. “Gender Bias in Neural Natural Language Processing.” Preprint, submitted July 31. https://doi.org/10.48550/arXiv.1807.11714.

Luccioni, Alexandra Sasha, Sylvain Viguier, and Anne-Laure Ligozat. 2022. “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model.” Preprint, submitted November 3. https://doi.org/10.48550/arxiv.2211.02001.

Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools.” [under review preprint] https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf.

Maiberg, Emanuel. 2023. “Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale.” *404 Media*, August 22. https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/.

Malevé, Nicolas, and Katrina Sluis. 2023. “The Photographic Pipeline of Machine Vision; or, Machine Vision’s Latent Photographic Theory.” *Critical AI* 1, nos. 1–2. https://doi.org/10.1215/2834703x-10734066.

Malik, Shreshth. 2021. “Deep Learning and Compute: Can We Just Keep Scaling?” *UCL Finance and Technology Review*, February 16. https://www.uclftr.com/post/ai-and-computing-power.

Mandaro, Laura, Natasha Mascarenhas, and Stephanie Palazzolo. 2024. “OpenAI Board Reappoints Altman and Adds Three Other Directors.” *Information*, March 9. https://www.theinformation.com/articles/sam-altman-to-return-to-openai-board-of-directors.

Mann, Jyoti. 2024. “Read the Memo Satya Nadella Sent to Employees Announcing Google

DeepMind cofounder’s Move to Microsoft.” March 20, https://www.businessinsider.com/satya-nadella-mustafa-suleyman-microsoft-memo-2024-3.

Manyika, James, and Michael Spence. 2023. “The Coming AI Economic Revolution: Can Artificial Intelligence Reverse the Productivity Slowdown?” *Foreign Affairs*, October 24. https://www.foreignaffairs.com/world/coming-ai-economic-revolution.

Marcus, Gary F. 2023. “The Sparks of AGI? Or the End of Science?” *Marcus on AI*, March 24. https://garymarcus.substack.com/p/the-sparks-of-agi-or-the-end-of-science.

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

Martin, Henri-Jean. 1994. *The History and Power of Writing*. Chicago: University of Chicago Press.

Marx, Paris. 2023. “How Sam Altman Plays into Microsoft’s Ambitions.” *Disconnect*, November 20. https://www.disconnect.blog/p/how-sam-altman-plays-into-microsofts.

McCarthy, J., M. L. Minsky, N. Rochester, C. E. Shannon. 1955. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.” John McCarthy (website), April 3, 1996. https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

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

McKinsey & Company. 2022. “What Are Industry 4.0, the Fourth Industrial Revolution, and 4IR?” August 17. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-industry-4-0-the-fourth-industrial-revolution-and-4ir.

Mercado, Angely. 2023. “Microsoft Is Using a Hell of a Lot of Water to Flood the World with AI.” *Gizmodo*, September 11. https://gizmodo.com/microsoft-water-usage-ai-iowa-data-center-1850826419.

Merchant, Brian. 2023a. *Blood in the Machine: The Origins of the Rebellion against Big Tech*. Boston: Little, Brown.

Merchant, Brian. 2023b. “Why Silicon Valley Hated Sam Altman’s OpenAI Ouster.” *Los Angeles Times*, November 21. https://www.latimes.com/business/technology/story/2023-11-20/column-openais-board-had-safety-concerns-big-tech-obliterated-them-in-48-hours.

Merchant, Brian. 2023c. “The Writers’ Strike Was a Victory for Humans over AI.” *Los Angeles Times*, October 9. https://www.latimes.com/business/technology/story/2023-09-25/column-sag-aftra-strike-writers-victory-humans-over-ai.

Messeri, Lisa and M. J. Crockett. 2024. “Artificial Intelligence and Illusions of Understanding in

Scientific Research.” *Nature* 627, no. 8002 (2024): 49-58.

Metz, Cade. 2023. “‘The Godfather of AI’ Leaves Google and Warns of Danger Ahead.” *New York Times*, May 4. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.’

Metz, Rachel. 2024. “OpenAI’s Co-Founder and Chief Scientist Ilya Sutskever Is Leaving the Company.” *Time,* May 16. https://time.com/6978195/ilya-sutskever-leaves-open-ai/.

Mickle, Tripp, Cade Metz, Mike Isaac, and Karen Weise. 2023. “Inside OpenAI’s Crisis over the Future of Artificial Intelligence.” *New York Times*, December 9. https://www.nytimes.com/2023/12/09/technology/openai-altman-inside-crisis.html.

Milmo, Dan, and Hibaq Farah. 2023. “Malicious Use of AI Could Cause ‘Unimaginable’ Damage, Says UN Boss.” *Guardian*, October 25. https://www.theguardian.com/technology/2023/jul/18/malicious-use-of-ai-could-cause-huge-damage-says-un-boss.

Mitchell, Margaret (@mmitchell_ai). 2023. “Finally have a moment to read MSR’s ‘Sparks of AGI’ paper.” Twitter thread, April 10, starting at 7:35 p.m. https://twitter.com/mmitchell_ai/status/1645571158585253888.

Mitchell, Melanie. 2023. “Can Large Language Models Reason?” *AI: A Guide for Thinking Humans*, September 10. https://aiguide.substack.com/p/can-large-language-models-reason.

Mok, Aaron. 2023. “ChatGPT Could Cost over $700,000 Per Day to Operate. Microsoft Is Reportedly Trying to Make It Cheaper.” *Business Insider*, April 20. https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4.

Monserrate, Steven Gonzalez. 2022. “The Staggering Ecological Impacts of Computation and the Cloud.” *MIT Press Reader*, February 14. https://thereader.mitpress.mit.edu/the-staggering-ecological-impacts-of-computation-and-the-cloud/.

Nagpaul, Sunny. 2024. “Four Worrying Trends in Tech That Are Fueling Google and Amazon Layoffs.” *Fortune*, January 20. https://fortune.com/2024/01/20/google-amazon-tech-layoffs-reasons-4-worrying-trends/.

Naughton, John. 2023. “Users, Advertisers—We Are All Trapped in the ‘Enshittification’ of the Internet.” *Guardian*, March 11. https://www.theguardian.com/commentisfree/2023/mar/11/users-advertisers-we-are-all-trapped-in-the-enshittification-of-the-internet.

NBC News. 2023. “Fmr. Google CEO Says No One in Government Can Get AI Regulation ‘Right.’” May 15. https://www.nbcnews.com/meet-the-press/video/fmr-google-ceo-says-no-one-in-government-can-get-ai-regulation-right-174442053869.

Newfield, Christopher. 2023. “How to Make ‘AI’ Intelligent; or, The Question of Epistemic Equality.” *Critical AI* 1, nos. 1–2. https://doi.org/10.1215/2834703x-10734076.

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

O’Brien, Matt, Hannah Fingerhut, and Associated Press. 2023. “A.I. Tools Fueled a 34% Spike in Microsoft’s Water Consumption, and One City with Its Data Centers Is Concerned about the Effect on Residential Supply.” *Fortune*, September 9. https://fortune.com/2023/09/09/ai-chatgpt-usage-fuels-spike-in-microsoft-water-consumption/.

Omiye, Jesutofunmi A., Jonathan Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. “Large Language Models Propagate Race-Based Medicine.” *npj Digital Medicine* 6, no. 1. https://doi.org/10.1038/s41746-023-00939-z.

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

O’Neil, Lorena. 2023. “These Women Tried to Warn Us about AI.” *Rolling Stone*, August 12. https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/.

Open AI. 2023a. “Frontier Model Forum.” Blog, July 26. https://openai.com/blog/frontier-model-forum.

Open AI. 2023b. “Frontier Risk and Preparedness.” Blog, October 26. https://openai.com/blog/frontier-risk-and-preparedness.

Open AI. 2024a. “OpenAI and Journalism.” Blog, January 8. https://openai.com/blog/openai-and-journalism.

Open AI. 2024b. “Hello GPT-4o.” Blog, May 13. https://openai.com/index/hello-gpt-4o/.

O'Sullivan, Donie, and Allison Gordon. 2023. "How Microsoft Is Making a Mess of the News after Replacing Staff with AI." *CNN*, November 2. https://edition.cnn.com/2023/11/02/tech/microsoft-ai-news/index.html.

Ouyang, Long, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." Preprint, submitted March 4. https://doi.org/10.48550/arXiv.2203.02155.

Pandey, Mohit, Michael Fernández, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C. Stern, and Artem Cherkasov. 2022. "The Transformational Role of GPU Computing and Deep Learning in Drug Discovery." *Nature Machine Intelligence* 4, no. 3: 211–21. https://doi.org/10.1038/s42256-022-00463-x.

Pasquale, Frank. 2020. *New Laws of Robotics: Defending Human Expertise in the Age of AI*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674250062.

Pasquinelli, Matteo. 2023. *The Eye of the Master: A Social History of Artificial Intelligence*. London: Verso Books.

Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Manguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. "Carbon Emissions and Large Neural Network Training." Preprint, last revised April 23. https://doi.org/10.48550/arXiv.2104.10350.

Pearl, Judea, and Dana Mackenzie. 2019. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Perrigo, Billy. 2023a. "Big Tech Is Already Lobbying to Water Down Europe's AI Rules." *Time*, April 21. https://time.com/6273694/ai-regulation-europe/.

Perrigo, Billy. 2023b. "Exclusive: 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/.

Pichai, Sundar. 2023. "An Important Next Step on Our AI Journey." *Keyword* (blog), Google, February 6. https://blog.google/technology/ai/bard-google-ai-search-updates/.

Piltch, Avram. 2023. "Google's AI Bots Tout 'Benefits' of Genocide, Slavery, Fascism, Other Evils." *Tom's Hardware*, August 22. https://www.tomshardware.com/news/google-bots-tout-slavery-genocide.

Raji, Inioluwa Deborah, Emily M. Bender, Amandalynne Paullada, Emily Denton, Alex Hanna. 2021. "AI and the Everything in the Whole Wide World Benchmark." Preprint, submitted November 26. https://doi.org/10.48550/arXiv.2111.15366.

Raji, Inioluwa Deborah, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. "Saving Face: Investigating the Ethical Concerns of

Facial Recognition Auditing." Preprint, submitted January 3. https://doi.org/10.48550/arXiv.2001.00964.

Ramel, David. 2024. "New GitHub Copilot Research Finds 'Downward Pressure on Code Quality.'" *Visual Studio Magazine*, January 25. https://visualstudiomagazine.com/articles/2024/01/25/copilot-research.aspx.

Rathi, Akshat and Dina Bass. 2024. "Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30%." *Bloomberg,* May 15. https://www.bloomberg.com/news/articles/2024-05-15/microsoft-s-ai-investment-imperils-climate-goal-as-emissions-jump-30?embedded-checkout=true.

Reisner, Alex. 2023. "Revealed: The Authors Whose Pirated Books Are Powering Generative AI." *Atlantic*, August 19. https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/.

Ross, Joel, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. "Who Are the Crowdworkers?" In *CHI EA '10: CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 2863–72. New York: Association for Computing Machinery. https://doi.org/10.1145/1753846.1753873.

Rotman, David. 2023. "ChatGPT Is about to Revolutionize the Economy. We Need to Decide What That Looks Like." *MIT Technology Review*, March 25. https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/.

Rushkoff, Douglas. 2022. *Survival of the Richest: Escape Fantasies of the Tech Billionaires*. New York: W. W. Norton.

Rushkoff, Douglas. 2023. "'We Will Coup Whoever We Want!': The Unbearable Hubris of Musk and the Billionaire Tech Bros." *Guardian*, November 28. https://www.theguardian.com/books/2023/nov/25/we-will-coup-whoever-we-want-the-unbearable-hubris-of-musk-and-the-billionaire-tech-bros.

Sadowski, Jathan. 2019. "When Data Is Capital: Datafication, Accumulation, and Extraction." *Big Data and Society* 6, no. 1. https://doi.org/10.1177/205395171882054.

Sahlins, Marshall. 1985. *Islands of History*. Chicago: University of Chicago Press.

Schaeffer, Rylan, Brando Miranda, Sanmi Koyejo. 2023. "Are Emergent Abilities of Large Language Models a Mirage?" Preprint, submitted April 28. https://arxiv.org/abs/2304.15004.

Schiff, Kaylyn Jackson, Daniel S. Schiff, and Natalia Bueno. 2022. "The Liar's Dividend: Can Politicians Use Deepfakes and Fake News to Evade Accountability?" Preprint, submitted May 10. https://doi.org/10.31235/osf.io/q6mwn.

Searle, John. R. 1980. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3, no. 3: 417–24. https://doi.org/10.1017/S0140525X00005756.

Shah, Chirag, and Emily M. Bender. 2024. “Envisioning Information Access Systems: What Makes for Good Tools and a Healthy Web?” *ACM Transactions on the Web*, February 26. https://doi.org/10.1145/3649468.

Sheng, Emily, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. “The Woman Worked as a Babysitter: On Biases in Language Generation.” In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the Ninth International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3407–12. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/d19-1339.

Shepard, Kenneth. 2024. “AI Art Generator Allegedly Scraped *Magic: The Gathering* Cards for Material.” *Kotaku*, January 5. https://kotaku.com/magic-the-gathering-midjourney-ai-art-lawsuit-1851143250.

Shieber, Stuart M. 1994. “Lessons from a Restricted Turing Test.” *Communications of the ACM* 37, no. 6: 70–78. https://doi.org/10.1145/175208.175217.

Shieber, Stuart M. 2004. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, MA: MIT Press.

Shieber, Stuart M. 2007. “The Turing Test as Interactive Proof.” *Noûs* 41, no. 4: 686–713. https://doi.org/10.1111/j.1468-0068.2007.00636.x.

Shift Project. 2019. “‘Lean ICT: Towards Digital Sobriety’: Our New Report on the Environmental Impact of ICT.” March 6. https://theshiftproject.org/en/article/lean-ict-our-new-report/.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. “The Curse of Recursion: Training on Generated Data Makes Models Forget.” *arXiv preprint arXiv:2305.17493* (2023).

Sigalos, MacKenzie, and Ryan Browne. 2024. “OpenAI’s Sam Altman Says Human-Level AI Is Coming but Will Change World Much Less than We Think.” *CNBC*, January 16. https://www.cnbc.com/2024/01/16/openais-sam-altman-agi-coming-but-is-less-impactful-than-we-think.html.

Skrentny, John David. 2024. “Why Pushing STEM Majors Is Turning Out to Be a Terrible Investment.” *Los Angeles Times*, January 9. https://www.latimes.com/opinion/story/2024-01-09/science-jobs-technology-stem-majors.

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

Sorkin, Andrew Ross, Ravi Mattu, Bernhard Warner, Sarah Kellser, Michael J. de la Merced, Lauren Hirsch, and Ephrat Livni. 2024. “The F.T.C. Takes on A.I. Deals” *New York Times*, January 26.

https://www.nytimes.com/2024/01/26/business/dealbook/ftc-ai-deals-microsoft-openai.html.

Spiers, Elizabeth. 2023. “A Tech Overlord’s Horrifying, Silly Vision for Who Should Rule the World.” *New York Times*, October 28. https://www.nytimes.com/2023/10/28/opinion/marc-andreessen-manifesto-techno-optimism.html.

Stark, Luke. 2023. “Breaking Up (with) AI Ethics.” *American Literature* 95, no. 2: 365–79. https://doi.org/10.1215/00029831-10575148.

Statt, Nick. 2016. “Microsoft Is Partnering with Elon Musk’s OpenAI to Protect Humanity’s Best Interests.” *Verge*, November 15. https://www.theverge.com/2016/11/15/13639904/microsoft-openai-ai-partnership-elon-musk-sam-altman.

Stoller, Matt. 2019. *Goliath: The One-Hundred-Year War between Monopoly Power and Democracy*. New York: Simon and Schuster.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. “Energy and Policy Considerations for Deep Learning in NLP.” In *Proceedings of the Fifty-Seventh Annual Meeting of the Association for Computational Linguistics*, 3645–50. Stroudsburg, PA: Association for Computational Linguistics.

Suchman, Lucy A. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. 2nd ed. Cambridge: Cambridge University Press.

Suleyman, Mustafa. 2023. *The Coming Wave: Technology, Power, and the Twenty-First Century’s Greatest Dilemma*. New York: Crown.

Tangermann, Victor. 2024a. “OpenAI Says It’s Fine to Vacuum Up Everyone’s Content and Charge for It without Paying Them.” *Futurism*, January 10. https://futurism.com/openai-content-new-york-times-lawsuit.

Tangermann, Victor. 2024b. “Sam Altman Says AI Using Too Much Energy, Will Require Breakthrough Energy Source.” *Futurism*, January 17. https://futurism.com/sam-altman-energy-breakthrough.

Tenbarge, Kat. 2023. “Found through Google, Bought with Visa and Mastercard: Inside the Deepfake Porn Economy.” *NBC News*, March 27. https://www.nbcnews.com/tech/internet/deepfake-porn-ai-mr-deep-fake-economy-google-visa-mastercard-download-rcna75071.

Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2020. “The Computational Limits of Deep Learning.” Preprint, last revised July 27. https://doi.org/10.48550/arXiv.2007.05558.

Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2021. “Deep Learning’s Diminishing Returns.” *IEEE Spectrum*, September 24. https://spectrum.ieee.org/deep-learning-computational-cost.

Thomson, Craig, Ehud Reiter. 2020. “A Gold Standard Methodology for Evaluating Accuracy in Data-to-Text Systems.” Preprint, submitted November 8. https://arxiv.org/abs/2011.03992.

Thorbecke, Catherine. 2023. “Elon Musk Announces a New AI Company.” *CNN*, July 12. https://www.wbaltv.com/article/elon-musk-ai-company/44519627.

Tiku, Natasha. 2024. “OpenAI Didn’t Copy Scarlett Johansson’s Voice for ChatGPT, Records Show.” May 23. https://www.washingtonpost.com/technology/2024/05/22/openai-scarlett-johansson-chatgpt-ai-voice/.

Tomasello, Michael. 1999. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tomasello, Michael. 2019. *Becoming Human: A Theory of Ontogeny*. Cambridge, MA: Harvard University Press.

Tso, Kathryn. 2023. “How Much Is a Ton of Carbon Dioxide?” *MIT Climate Portal*, December 21. https://climate.mit.edu/ask-mit/how-much-ton-carbon-dioxide.

Turing, Alan. 1950. “Computing Machinery and Intelligence.” *Mind*, no. 236: 433–60. https://doi.org/10.1093/mind/lix.236.433.

Upwork Team. 2023. “Ten Benefits of Generative AI: Increase Productivity and Creativity.” Blog, September 20. https://www.upwork.com/resources/generative-ai-benefits.

US Office of Science and Technology Policy. 2022. “Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People.” October. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

van Rooij, Iris. 2022. “Against Automated Plagiarism.” Blog, December 29. https://irisvanrooijcogsci.com/2022/12/29/against-automated-plagiarism/.

Verma, Pranshu, Nitasha Tiku, and Gerrit De Vynck. 2023. “Sam Altman Reinstated as OpenAI CEO with New Board Members.” *Washington Post*, November 22. https://www.washingtonpost.com/technology/2023/11/22/sam-altman-back-openai/.

Vincent, James. 2023. “A Data Scientist Cloned His Best Friends’ Group Chat Using AI.” *Verge*, April 13. https://www.theverge.com/2023/4/13/23671059/ai-chatbot-clone-group-chat.

Wan, Yixin, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. “‘Kelly Is a Warm Person, Joseph Is a Role Model’: Gender Biases in LLM-Generated Reference Letters.” Preprint, submitted October 13. https://doi.org/10.48550/arXiv.2310.09219.

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2024. “Large Language Models Cannot Replace Human Participants Because They Cannot Portray Identity Groups.” *arXiv preprint arXiv:2402.01908*.

Wang, Boxin, et al. 2023. “DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models.” Preprint, submitted June 20. https://doi.org/10.48550/arXiv.2306.11698.

Warschauer, Mark. 2010. “New Tools for Teaching Writing.” *Language Learning and Technology* 14, no. 1: 3–8. https://www.lltjournal.org/item/10125-44196/.

Watters, Audrey. 2023. *Teaching Machines: The History of Personalized Learning*. Cambridge, MA: MIT Press.

Weatherbed, Jess. 2023. “The *New York Times* Prohibits Using Its Content to Train AI Models.” *Verge*, August 14. https://www.theverge.com/2023/8/14/23831109/the-new-york-times-ai-web-scraping-rules-terms-of-service.

Weidinger, Laura, et al. 2021. “Ethical and Social Risks of Harm from Language Models.” Preprint, submitted December 8. https://doi.org/10.48550/arXiv.2112.04359.

Weiser, Benjamin. 2023. “A Man Sued Avianca Airline. His Lawyer Used ChatGPT.” *New York Times*, May 27. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

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

Weizenbaum, Joseph. 1966. “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine.” *Communications of the ACM* 9, no. 1: 36–45.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. London: W. H. Freeman.

West, Sarah Myers. 2023. “AI Now’s Testimony to the US Congress over Algorithms and Competition.” AI Now Institute, December 14. https://ainowinstitute.org/publication/ai-nows-testimony-to-the-us-senate-on-algorithms-and-competition.

Wheatley, Mike. 2024. “Google CEO Sundar Pichai Warns More Job Cuts Will Be Necessary to Achieve ‘Ambitious Goals’ This Year.” *SiliconANGLE*, January 19. https://siliconangle.com/2024/01/18/google-ceo-sundar-pichai-warns-job-cuts-will-necessary-achieve-ambitious-goals-year/.

Whittaker, Meredith. 2021. “The Steep Cost of Capture.” *Interactions* 28, no. 6: 50–55. https://doi.org/10.1145/3488666.

Widder, David Gray. 2024. “Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints.” *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24),* June 3–6, 2024, Rio de Janeiro, Brazil.ACM, https://arxiv.org/pdf/2402.08171.

Wiggins, Kyle. 2024. “AI Training Data Has a Price Tag That Only Big Tech Can Afford.”

TechCrunch, June 1. https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford/.

Williams, Raymond. 1985. *Keywords: A Vocabulary of Culture and Society*. New York: Routledge.

Wong, Matteo. 2023. "Humans Are Haunting the Chatbots." *Atlantic*, August 11. https://www.theatlantic.com/technology/archive/2023/07/ai-chatbot-human-evaluator-feedback/674805/.

Zakrzewski, Cat. 2023. "FTC Investigates OpenAI over Data Leak and ChatGPT's Inaccuracy." *Washington Post*, July 13. https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/.

Zuboff, Shoshana. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

Figure 1. Lauren Goodlad's September 2022 query of GPT 3.5: prompt appears in brown; yellow highlights indicate false information. **{Au: Please confirm changes to figure numbers CONFIRMED**

Figure 2. November 2023 query of Bing Chat (harnessed to GPT-4).