Uncorrected preprint of
Bender, Emily M, 'Artificial Intelligence' (25 June 2026), in Rayvon Fouché (ed.), Oxford Research Encyclopedia of Science, Technology, and Society (New York, NY, online edn, Oxford Academic, 25 June 2026 -), https://doi.org/10.1093/9780197852712.003.0067

Artificial Intelligence

Emily M. Bender

January 14, 2026

Article Summary

The term *artificial intelligence* is typically used as if it refers to a coherent, extant or near-future set of technologies, but in fact it does not. This article traces the history of the notion of artificial intelligence and the various ways that the idea has been used to organize how we understand our world, allocate resources, and relate to each other.

Keywords

Research practice, technology hype, technology and power, labor, Eliza effect

1 Introduction

The term *artificial intelligence (AI)* was coined by John McCarthy in a grant proposal, written jointly with Marvin Minsky, Nathaniel Rochester and Claude E. Shannon for a two-month long summer project held at Dartmouth in 1956 (McCarthy et al., 1955). Even in this initial use, it was a marketing term: McCarthy et al needed a wrapper that would tie together the diverse problems they hoped to work on and appeal to their funding agency, while distancing themselves from adjacent work by Norbert Wiener and colleagues that was being carried out under the name *cybernetics*.

The idea of thinking machines is older, though, with many antecedents in fiction. Early fictional treatments include Karel Čapek's 1920 play R.U.R, the source of the word *robot*, in which the robots were artificial biological

entities; Murray Leinster's 1946 short story "A Logic Named Joe" in which the *logic* (a computer) gains sentience, and Isaac Asimov's robot stories, including the 1942 short story "Runaround", in which he first presents his Three Laws of Robotics.

This literary history is important because of the way it shapes the public's imagination. When scientists and technologists earnestly speak of artificial intelligence or thinking machines, they leverage their audiences' familiarity with the speculative worlds of science fiction.

The term *artificial intelligence* does not refer to a coherent set of technologies and it never has. In 1956, McCarthy et al proposed to work on programming languages, computational complexity, artificial neurons and neural networks, "self-improvement" by machines, and other research tasks. Their pitch for this project included 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" (McCarthy et al., 1955, p.2). By labeling this a conjecture, McCarthy et al avoid needing to support it or precisely define the terms involved and instead are able to continue their argumentation based on the unfounded assumption that 'learning' and 'intelligence' are computable.

In 2025, the term *artificial intelligence* continues to be applied to diverse and unrelated technologies while still lacking any precise definition. Technology that falls under the AI umbrella includes systems for calculating the likely shape of a protein given a sequence of amino acids (e.g. Jumper et al., 2021), systems for calculating good next moves in games like chess (e.g. McCarthy, 1990) or Go (e.g. Silver et al., 2017), machine translation software (e.g. Brown et al., 1993), automatic transcription software (e.g. Davis et al., 1952), recommendation systems (e.g. Goldberg et al., 1992), systems for matching photographs of the same person (e.g. Taigman et al., 2014), synthetic media generating machines (creating images, video, and text, in the latter case called large language models or LLMs) (e.g. Brown et al., 2020) and many others, including systems based explicitly in modern-day physiognomy and purporting to do things like predict whether a person is a 'criminal' based on a photo of their face (see Stark and Hutson, 2022; Agüera y Arcas et al., 2023).

The lumping together of these diverse technologies under a single umbrella is part of the marketing pitch, inasmuch as it suggests that we have one piece of technology that can handle protein folding, playing chess, translating between languages, and respond with apparently coherent text to any input

query. Because these functions are so diverse, if it is one thing doing all of them, that creates the illusion of something that can learn flexibly. But in fact, these are all separate systems, and even if some of them use similar components, they are all purpose-built.

The term *artificial intelligence* resists definition because it is continually reappropriated by people to mean different things. This, in turn, means that discussions of artificial intelligence that don't provide working definitions for the purposes at hand risk incoherence. Attempts at definition that aren't focused on specific technologies under discussion tend to fall apart because they rely on the idea of 'intelligence' in humans (an ill-defined concept, based on race science and the idea that there is a single number that can represent an individual's cognitive abilities (Gould, 1981)) or vague appeals to similarity to humans. Is it 'AI' if the system can do something that (previously) only people could do? Or is it 'AI' only if the system can do anything people could do? (And what counts as something people can do?)

Accordingly, this article does not provide a definition of the term artificial intelligence, but rather explores various ways in which the idea of artificial intelligence has been used to organize how we understand our world, allocate resources, and relate to each other.

2 AI is a research field

The phrase artificial intelligence names a research field, sometimes understood as a subfield of computer science and other times as an interdisciplinary endeavor also involving other fields such as electrical engineering, cognitive science, psychology, neuroscience, linguistics and philosophy. This research field has its own conferences and journals. Prominent conferences include IJCAI (the International Joint Conferences on Artificial Intelligence, since 1969), AAAI (the Association for the Advancement of Artificial Intelligence, since 1980), ICML (the International Conference on Machine Learning, since 1980), NeurIPS (Neural Information Processing Systems, since 1987) and ICLR (the International Conference on Learning Representations, since 2013). Prominent journals include *Artificial Intelligence Review* (since 1987) and the *Journal of Machine Learning Research* (since 2000).

As a research field, artificial intelligence has had phases of being intensely popular and phases of being deeply out of fashion. These latter phases are called 'AI winters', a moniker that takes the perspective of proponents of the

field and laments the lack of funding (i.e. sunlight or input energy) for it. These periods of lack of funding are understood to be the result of earlier hype cycles in which AI researchers overpromised what their technology could do, leading to disillusionment on the part of funders. Early AI research was largely funded by military and national intelligence agency research contracts and the first burst of funding was cut back significantly in the US following the 1966 ALPAC report (ALPAC, 1966) and in the UK following the 1973 Lighthill report (Lighthill, 1973), both of which referred to great enthusiasm but lackluster results.

This research field can be understood as focused on a collection of specific tasks or problems, such as image and video processing, text and speech processing, knowledge base creation, planning, theorem proving, etc. Each of these areas has its own research community, with specialized conferences and journals. During 'AI winters', when 'AI' as an overarching research goal is out of fashion, research focused more specifically on these practical goals continues, often without the 'AI' moniker. Across all of these tasks, there have been both symbolic or rules-based and statistical or probabilistic approaches tried, with the statistical/probabilistic approaches (also called machine learning) gaining in popularity over time. Statistical/probabilistic approaches involve designing algorithms that can extract patterns from input datasets ('training data') and then use those patterns to provide outputs for additional inputs not included in the training data.

The research field of AI can also be understood as a search for algorithmic approaches which generalize across these different specific tasks. Indeed, Birhane et al. (2022) find that 'generalization' is the second most widely referenced value in the collection of machine learning papers they studied, sampled from 2008, 2009, 2018 and 2019. Generalization itself can be seen in a few ways (Blili-Hamelin et al., 2025): In the first instance, it is a measure of how well a system works beyond the specific data or scenarios it was developed with. This version of generalization underlies the standard practice of evaluating systems on 'held-out' data, or data that was not used in system development. A second notion of generalization is the idea that a specific type of algorithm can be applied to different types of training data, making the algorithm 'general purpose' across different tasks, though any given system trained on a specific task is specialized for that task. A third, and far more problematic, notion of generalization is one in which researchers are striving to create a 'general learning algorithm' that could learn as flexibly as people do.

The push for this third kind of generality leads to intertwined problems with both the scientific validity of research and negative societal impacts. On the one hand, sheer scale of dataset (e.g. with ImageNet; Deng et al. 2009) or claimed variety of subtasks (e.g. with SuperGLUE; Wang et al. 2019) have been sold as and then taken up as suitable proxies for 'generality'. This has led to wide-spread, unsupported claims of achieving general approaches to such tasks as 'visual understanding' or 'natural language understanding' (Raji et al., 2021). On the other hand, the push for generality, combined with the approach of designing algorithms to extract patterns from data, leads to a general datafication of experience (Couldry and Mejias, 2019). If machine learning is to be as general as human learning, and machines learn from data, then everything about human experience must be representable by data. This lens on the world leads to varied harms including exploitation of data workers (Fort et al., 2011; Hao, 2025), misconstrual of data about people as representative of the people themselves (Raji, 2020), and the legitimization of ill-founded tasks (such as detecting 'criminality', political stance or sexuality from images of people's faces) simply because a dataset sold as representing that task can be 'solved' at better-than-chance rates by machine learning algorithms (Paullada et al., 2021).

3 AI is an approach to cognitive science

In addition to being pursued for practical ends, work in artificial intelligence has also been motivated by scientific goals, specifically within cognitive science and psychology. Van Rooij et al. (2024) trace the history of the idea that building a computational model of one or more human cognitive systems (or ultimately an integrated model of all human cognitive systems, should these be enumerable) would be beneficial to the project of understanding those systems scientifically. After all, if we can build something, it stands to reason that we can understand what we built.[1]

[1]In this light, the discussions by LLM-promoters about the source of the 'capabilities' of their systems being mysterious, murky, or not understood (e.g. Sam Bowman in Hassenfeld, 2023) are quite suspect. In fact, the claims of 'capabilities' are based on extrapolation from performance on specific, finite evaluation sets ('benchmarks'). Without a theory of how the system as built produces expected outputs better than a random-chance baseline, those claims remain unsupported (Raji et al., 2021; Birhane et al., 2022). In the case of LLMs, an alternative hypothesis to claims of e.g., 'reasoning' capabilities is that there is a mismatch between what the system is doing (modeling the distribution of word

Van Rooij et al (2024) show that interest in using approaches to artificial intelligence as a means of understanding human cognition has waxed and waned over the years, with practitioners alternately wary of overinterpreting models as precise reflections of human cognition and enthusiastic about constructing artificial cognizers. We can find warnings against hype around computing in general and about the influence of computational thinking on science already in the work of Ada Lovelace herself, who wrote in personal correspondence (July 1843; cited in Toole 1998):

> The Analytical Engine has no pretensions whatever to *originate* any thing. It can do whatever we *know how to order it* to perform. It can *follow* analysis; but is has no power of *anticipating* any analytical relations or truths. Its province is to assist us in making *available* what we are already aquainted with. This it is calculated to effect primarily and chiefly of course, through its executive faculties; but it is likely to exert an indirect and reciprocal influence on science itself in another manner. [...] It is however pretty evident, on general principles, that in devising for mathematical truths a new form in which to record and throw themselves out for actual use, views are likely to be induced, which should again react on the more theoretical phase of the subject. (p.191–192; emphasis original)

Looking at the impact of computational models on modern cognitive science, Van Rooij et al (2024) distinguish between what they term *makeism* and and other approaches to *computationalism* which they argue better advance the goals of cognitive science. Makeism corresponds to a set of beliefs, widely evinced in work on artificial intelligence, that hold that:

> (a) it is possible to (re)make cognition computationally; (b) if we (re)make cognition, then we can explain and/or understand it; and possibly (c) explaining and/or understanding cognition requires (re)making cognition itself. (p.626)

This set of beliefs is evident at the founding of the field of AI, in McCarthy et al's conjecture, quoted above, about learning being precisely describable

forms in text, and being used to repeatedly output likely next words) and what the people ostensibly testing it are interpreting this as ('reasoning') (see also Bowers et al. (2023)).

enough for machine simulation. It continues in the work of Feigenbaum (1963) who writes:

> Researchers in the field [of artificial intelligence] hold to the working hypothesis that human thinking is wholly information-processing activity within the human nervous system; that ultimately, these information processes are perfectly explicable; [. . .] and that digital computers, being general information processing devices, can be programmed to carry out any and all of the information processes thus explicated. (p.1962–1963)

And it continues into more modern times. AI researcher Gary Marcus, writing in the *New York Times* asserts "If the heart is a biological pump, and the nose is a biological filter, the brain is a biological computer, a machine for processing information in lawful and systematic ways" (2015). This begs the question of whether the brain is in fact best understood as processing information, but Marcus is not alone hewing to this analogy in modern times. In 2022, OpenAI's Sam Altman, appropriating the coinage of Bender, Gebru et al (2021), tweeted "I am a stochastic parrot and so r u" (2022). Altman here is apparently attempting to diffuse a recharacterization of OpenAI's chatbot that undermines his claims to be creating artificial intelligence via text synthesis by asserting that in fact people also engage in similar processes. As a final example, take Anthropic's CEO Dario Amodei's imagining of "a country of geniuses in a data center" (2024), referring to AI programs that can think like people, but better.

This view of human cognitive processes as information processing, where information processing is defined in terms of what computers can do, is by now deeply ingrained (though not universally accepted) in both scientific and popular culture. Baria and Cross (2021) critique the use of what they term the *computational metaphor* ('The brain is a computer') in neuroscience. Though debated, rather than universally accepted in the field, this metaphor is nonetheless pervasive. For lay people not directly involved in neuroscience, its pervasiveness in popular culture discourse might leave the impression that it's not a metaphor but simply a truth of the world: We process information, computers process information, surely at some appropriate level of abstraction brains and computers must be the same thing.

It is instructive, however, to look to historical work on how people of previous eras have understood their own bodies via metaphors of science and

technology of the day. Smith (1993) provides a tour of various metaphors what have been used in work in psychology and brain science over the centuries, from Descartes describing the nervous system in terms of hydraulic powered machinery (17th century), through John Locke and David Hartley reaching for Newtonian physics as sources of metaphor (17th-18th century), to Herbert Spencer using a metaphor of the pianoforte, with different chords produced by striking different keys together, and piano mècanique, with "tune boards" in the grey matter which produce different "nerve currents" (19th century), to John Hughlings Jackson speaking about the brain as a hierarchical system like the military and into the 20th century fascination with telecommunications wiring, which John Z. Young uses as a metaphor and which is a clear predecessor to using computer chips in that role.[2] Smith writes about the computational metaphor seeming to reach the end of its life and popularity, with authors like Bergland (1980s) proposing that the biochemistry of glands is a better metaphor, though Smith also sees some resurgence of the computational metaphor in work in connectionism (aka "neural networks") and parallel and distributed processing. Undoubtedly, progress in biology, microscopy and other sciences have also had an impact on the metaphors chosen, but seeing this history helps locate our present use of the computational metaphor as a social fact, rather than a necessary one.

Van Rooij et al (2024), while arguing against makeism, nonetheless make the case that cognitive science should reclaim AI and what they term *computationalism* as a research tool (as in computational cognitive modeling; Guest and Martin, 2021; van Rooij and Baggio, 2021). In particular, they promote studying cognitive capacities in terms of computational problems, crucially without conflating the model (a computational implementation) with the system being modeled (some system within actual human cognition), because computationalism provides some useful scaffolds for theory building: On the one hand, it can be helpful to distinguish between levels of explanation (cognitive processes, capacities, and their physical implementations), which in turn help scientists align types of arguments or experiments with the level of explanation they actually contribute to. On the other hand, computationalism allows us to reason about tractability (van Rooij et al., 2019). A theory of cognitive processes or capacities can be tested for initial plausibility by asking if the processes/capacities posited are computable in principle.

In sum, AI has been and can be seen as an approach to studying cognitive

[2]For further historical detail, see also Marshall (1977).

science. However, van Rooij et al. (2024) are clear that progress in cognitive science is hampered by makeism, i.e. the approach to studying human cognition through attempting to engineer a replica of it. Thus for an academic field of AI-based cognitive science to succeed it would have to purse computationalism without makeism. Given that so much of the marketing around corporate work in AI relies on makeist assertions of meeting or exceeding human performance (e.g. He et al., 2021; Jamali and McMahon, 2025; Browne, 2025), however, it remains unclear at this time whether such a reclaiming of 'AI' by non-makeist cognitive scientists is possible.

4 AI is a parlor trick

Natural language processing, including natural language interfaces to computers, has been part of both the vision of artificial intelligence and the work that has taken place under that umbrella since the founding of the field or even before.[3] In a 1950 essay, Alan Turing proposed the "imitation game" (now called the Turing Test). He began with the question "Can machines think?" and rejected that question on the grounds that it would require definitions of both 'machine' and 'think'. Nonetheless, he asserts that his imitation game provides a problem which is "closely related" to it (1950, p.433). In the imitation game, an interrogator corresponds with two participants via type-written messages. One of those participants is a man who is attempting to help the interrogator, the other is pretending to be a man and is attempting to fool the interrogator. The interrogator has the job of determining which participant is actually the man. In one version of the game, the second participant is a woman. In another, the second participant is a machine, i.e. a computer. Turing asserts that we can then replace the question "Can machines think?" with the questions: "Will the interrogator decide wrongly as often when the game is played [with a computer] as he does when the game is played between a man and a woman?" (*Ibid.*:434)

It is easy to see how, in 1950, evidence that a computer that could convincingly carry out an open-ended conversation in English (or any other natural language) would have seemed like a very good proxy for evidence that the machine could actually 'think' in any ordinary sense of that word.

[3]Though, importantly, there is also robust work in natural language proessing/computational linguistics that is neither sold as nor motivated by the idea of artificial intelligence.

In fact, work on natural language processing frequently runs into ceiling effects where what is missing is a general model of the world and the ability to reason about it.[4]

However, what is required by the Turing Test is not an accurate representation of communicative intent as expressed in language, but rather the ability to simulate conversation, and this became apparent as early as 1966, with Joseph Weizenbaum's publication of his chatbot program ELIZA (Weizenbaum, 1966). At the time ground-breaking, this relatively simple text processing program was set up to output grammatical responses to user input in English, based on a simple set of rules. Some of the rules swapped first-person with second-person pronouns and made questions in response to statements by adding stock phrases (e.g. *You hate me.* ⇒ *What makes you think I hate you?*). Others triggered more apparently varied responses by providing completely separate sentences triggered by particular keywords (e.g. if the user's input included the word *everybody*, the system could say *Who in particular are you thinking of?*.) In the most well-known version of this program (dubbed DOCTOR, but now called just ELIZA), Weizenbaum finessed the problem of modeling the kind of world knowledge usually required to produce coherent conversation by choosing a speech situation that barely requires any: DOCTOR simulates a conversation with a Rogerian psychotherapist, where the user would expect to see their statements repeatedly reframed into questions, without added information.

Weizenbaum expected that anyone who saw the rules behind the program would understand that it was not actually understanding, but rather just responding mechanically. He was shocked to discover that people related to it quite differently (Weizenbaum, 1976). He describes both his secretary (unfortunately unnamed) asking him to leave the room so as to not see her private

[4]Take for example, the issue of gender on pronouns in machine translation output. Because gender systems aren't necessarily parallel across language pairs, what is required is resolution of the referent of the pronoun so that the appropriate target-language gender can be applied. But reference resolution, while approximable via surface textual patterns only, in the general case requires understanding the meaning of the text as well as social knowledge about the world being described. This is well illustrated in Terry Wingorad's famous example of the English sentences *The police denied the protestors a permit because they feared/advocated violence* (Winograd, 1972). Correct translation of these examples into French requires coreference resolution: In the first version, *they* most likely refers to the police, and so should be translated as *elle*; in the second it most likely refers to the protestors, giving *ils*, *elles*, *iels* depending on what is known about the gender of the protestors and whether the system is using inclusive language.

conversations with the machine, as well as professional psychologists excitedly talking about using the program to widen access to psychotherapy.[5] This has been dubbed the ELZA effect, which can be understood as the very human tendency to attribute understanding, thinking, or other types of cognition to computer systems that flexibly and fluently output natural language. The large language models behind chatbots like OpenAI's ChatGPT, Anthropic's Claude, Google's Gemini and others are more complex systems, where linguistic information is encoded not via rules but rather processing of large amounts of input ('training') text, but at a fundamental level they are also only manipulating linguistic form — and via the ELIZA effect, user experience of the system.

Research in linguistics in the intervening years, especially psycholinguistics and pragmatics (Reddy, 1979; Clark, 1996; Levinson, 2025), has provided insight that can help us understand why this might be. It is common to believe that when we understand spoken, written or signed language, we are unpacking meaning in a sequence of words that was packed into those words by their author. However, this is not the case. Instead, we are keeping in mind everything we know or believe about the author's knowledge and beliefs, about the common ground we share with the author, and about what the author likely knows or believes about their audience (be that us or some third party). Against that background, we then answer the question: What must they have been trying to convey by choosing those words with that grammatical structure? In other words, to make sense of language, we have to imagine a mind behind the text. This works just fine when the language was written by people or even a group of people, and it works reasonably well even when our assumptions about the author's beliefs are faulty. But it causes trouble when we encounter seemingly fluent and coherent text output by a mechanical process, because we still reflexively and instinctively apply the same process to make sense of it, imagining a mind that isn't there. Far from providing clear information about system affordances, the chat interface (turn-taking conversations; the use of first-person pronouns) common to ELIZA and today's chatbots leans into and enhances this illusion.

As of 2025, this parlour trick has been extremely persuasive and effective for driving media interest in large language models (and by extension other

[5]These experiences, among others, motivated Weizenbaum to become a life-long critic of the projects of artificial intelligence, automation and computerization in general (Tarnoff, 2023).

things called artificial intelligence) and effective at driving venture capital investment. Maslej et al. (2025) report that global investment in AI was $252.3 billion in 2024, of which $150.79 billion was private investment, a record high for that category and a 26% increase on 2023. 64% of funding to startups in the US in the first half of 2025 was to AI startups (Hu and Nishant, 2025). Not all systems sold as 'AI' involve LLMs, but the lumping together of diverse technologies, including LLMs/chatbots under the moniker 'AI' allows salespeople to borrow the effect of the parlour trick across diverse technologies.

The parlor trick is effective despite lack of evidence of reliability of these systems. Raji et al. (2022) find that the vast majority of work around 'AI', including studies of ethical deployment and proposed or implemented regulations, rest on the assumption that 'AI' systems function as claimed. They further find that this assumption is not generally supported. Surveying a database of documented failure cases, they find that machine learning-based and other 'AI' systems can fail at any number of levels including: (1) right at the conceptualization, being designed to approach tasks that are either conceptually or practically impossible; (2) in engineering, through failures of model design, failures of model implementation, or lack of safety features; and (3) in deployment, where the system might not be robust to ordinary usage (and this lack of robustness hidden by poor evaluation techniques), to adversarial attacks, or to unanticipated moves by end users. All of these failure modes are made more dangerous by the fourth category in Raji et al's taxonomy, failures of communication, wherein those selling 'AI' systems may falsify, overstate or misrepresent system functionality. Just as Bowers et al. (2023) call for in the use of AI models for cognitive science research, it is vitally important that any 'AI' systems (or systems not sold as 'AI' but built with machine learning) be rigorously and even 'severely' tested before use.

In summary, AI is a parlor trick (or 'con'; Bender and Hanna, 2025) on multiple levels: At the base level, the development of systems that mimic the way people use languages plays on the way we perceive and process language to create the illusion of a thinking entity. Because systems trained on enormous amounts of text can output plausible looking imitations of language use across many domains, researchers and companies can market such systems as nearly-there solutions in many different domains. Meanwhile, lumping disparate technologies together under the moniker of 'AI' facilitates the use of the illusion from language technology to bolster the credibility

of other kinds of systems. Finally, the present-day culture of 'AI' research facilitates unsupported claims by failing to insist on rigor in evaluation and scientific methodology more generally (Raji et al., 2022; Hullman et al., 2022; Blili-Hamelin et al., 2025).

5 AI is an ideology

In a short essay, Alkhatib (2024) argues that the best way to understand what the term artificial intelligence refers to is to understand it as an ideology. In particular, he concludes that "AI is an ideological project to shift authority and autonomy away from individuals, towards centralized structures of power."

In coming up with this definition, he pushes off from the examples that Narayanan and Kapoor (2024) provide to clarify their criteria for determining whether something is AI ("[D]oes the task require creative effort or training for a human to perform?" [p.12], "Was the behavior of the system directly specified in code by the developer, or did it indirectly emerge, say by learning from examples or searching through a database?" [p.13], and "[Does] the system [make] decisions more or less autonomously and [possess] some degree of flexibility and adaptability to the environment?" [p.13])

For Alkhatib, this location of the definition of AI in the technical details of the systems rather than the way they are conceived and used in the world misses the point. Where Narayanan and Kapoor want to define AI in terms of how systems are built, and in particular, whether there is any aspect of the system as deployed that coders cannot directly take credit for, Alkhatib directs our attention to the ways in which those claims of "autonomy" support disempowerment of the people who have to deal with the systems, especially as data subjects: "We can recognize, based on our own knowledge and experience as people who deal with these systems, what's part of this overarching project of disempowerment by the way that it renders autonomy farther away from us, by the way that it alienates our authority on the subjects of our own expertise" (Alkhatib, 2024).

This shift away from designer intent towards system impact is reminiscent of the cybernetician Beer's 2002 dictum that "The purpose of a system is what it does" (p.217). In the case of AI systems, McQuillan (2022) warns of "resonances between AI and the emergence of fascistic solutions to social problems" (p.5). He points to the ways in which systems called AI are built

on "pervasive data surveillance and centralized control" (p.14) and through the optimization operations embedded "a kind of abstract utilitarianism" (p.15).

Gebru and Torres (2024) situate the goal of artificial general intelligence (as opposed to narrow AI applications that target specific tasks) within an ideological tradition they dub the "TESCREAL bundle". This acronym stands for Transhumanism, Extropianism, Singulatarianism, Rationalism, Effective Altruism and Longtermism. Through close examination of primary texts, Gebru and Torres trace the lineages connecting all of these ideologies, including overlapping communities of practice, and their roots in 20th century eugenics as promoted and practiced in the US and the UK.[6] In these ideologies, people are ranked according to their (genetically-endowed) 'intelligence', and the goal is to variously produce an 'AGI' that supersedes even the highest-ranked people in that scheme or to produce 'enhanced' people through human-machine mergers. Again taking an extremely utilitarian point of view, the intrinsic value of people, especially those not ranked high on this imagined 'intelligence' scale, is disregarded and any present-day suffering counted as immaterial against a far future with extremely large numbers of simulated people living in data centers across the galaxy.

Several scholars have surfaced and analyzed the ways in which data-driven automated systems (often, though not always, called 'AI') serve to implement ideologies of racism and others forms of oppression. Sweeney (2013) investigates how online ad delivery systems, serving the interests of advertisers and the companies selling ad space, produce discriminatory outcomes. In the particular case she documents, the harm is the pervasive suggestion of criminal history associated to African-American-coded names. Noble (2018) follows this up with a deep exploration of how Google's commercialization of information access results in the commodification of identity terms and sets the stage for searches like "Black girls" to return pages of links to porn sites all while creating the impression that this is just how the world (or at least the World Wide Web) is.

Abdurahman (2022) describes pervasive datafication as families interact with social services and the child welfare system (more aptly known as the family policing system, per Abdurahman). This process serves as one step in a chain of actions that allows this system to separate families, not least

[6]For a connection between the view of brains as computers, as critiqued in Baria and Cross 2021 and eugeinics, see Benjamin 2024.

because the data is not grounded in the actual experience of families and is designed to disguise abuse on the part of the system and its component agencies. The datafication process, including information about accessing state-provided but not private mental health services, feeds machine learning systems like Allegheny County, PA's Allegheny Family Screening Tool, and constitutes what Eubanks (2018) calls the "digital poorhouse".

Raji (2020) identifies the reverberations of white supremacy in the way that datasets for machine learning erase, other, criminalize and/or pathologize Black people especially. She goes on to describe the danger that ensues when these datasets, constructed according to the lies of white supremacy, are treated as 'ground truth' for the creation and evaluation of machine learning/AI systems.

To understand artificial intelligence as an ideology is to look at it as a way of organizing ourselves and our world. In particular, artificial intelligence as an ideology is an ideology of oppression and control rather than liberation (McQuillan, 2022; Alkhatib, 2024), as illustrated by the examples cited above: The idea that decisions of any kind impacting people could be made autonomously by machines, away from the *metis* of people (Scott, 1998) who are accountable for the decisions, necessitates that people be treated as datafied subjects and both supports and relies on such an ideology.

6 AI is a way to hide and devalue human labor

Companies and researchers selling AI products are incentivized to have their audiences (consumers, other researchers, the broader public) to believe that the functionality of the systems inheres in either the 'intelligence' of systems themselves or the engineering work that is done to produce them. However, as documented by the research firm Cognilytica (2020), approximately 80% of the effort of creating a machine-learning based product is data work. As Bender and Hanna (2025, p.58) put it, "AI is always people."

Data work is intensely devalued. On the one hand, there are multiple different ways in which producers of AI systems attempt to appropriate the labor of people without any compensation. The training data for the large language models of the early 2020s (OpenAI's GPT series, Facebook's Llama series, Anthropic's Claude series, etc.) is comprised in large part of data

scraped from the internet as well as libraries of pirated books (Jia and Nagaraj, 2025). The training data for image generation systems like DALL-E, Midjourney, and Stable Diffusion similarly consists of appropriated artwork (Chayka, 2023).

This view of the result of people's effort being free for the taking has precedents earlier in the century in the work of von Ahn and others on what they termed "games with a purpose" and "CAPTCHAs". Writing in *Communications of the ACM*, von Ahn and Dabbish (2008) observed that people spend a lot of time playing games and asked "What if this time and energy were also channeled toward solving computational problems and training AI algorithms?" (p.60) and also talk about "harnessing human processing skills through computer games" (*Ibid.*, p.60). The idea behind games with a purpose is that game designers create games that people will play willingly and derive enjoyment from, while producing data labels that are of use as training data. A similar idea shows up in Chatbot Arena (Chiang et al., 2024), billed as a "platform for evaluating LLMs by human preference" (p.1) that overcomes issues with static benchmarks. The purported value to the users, who provide their testing and rating labor for free, is the opportunity to "easily access, explore and interact the world's leading AI models" (LMArena, 2025).[7]

CAPTCHAs or "Completely Automated Turing test to tell Computers and Humans Apart" (von Ahn et al., 2003), on the other hand, allow data collectors to take advantage of micro increments of labor required of people trying to access web pages. This is presented to the public as a beneficial system for combatting spam/automated attacks on web pages, but sold to industry as a way to create training data. Pettis (2023) argues that CAPTCHAs serve the purposes of reducing spam, building training data, and defining norms of what it means to be a user of web technology, all three of which priamrily serve industry's goals, rather than users'. He describes a 2006 talk by Luis von Ahn to Google employees in which von Ahn "describes his concern that any amount of time, no matter how small, that a human being is not being productive is wasted time, and represents valuable human cycles that can be used to solve open problems of computation" (Pettis, 2023, p.892). Google has used CAPTCHAs to produce training data for such things as optical character recognition (OCR) systems and image processing

[7]That chatbots are not in fact suitable technology for information access (Shah and Bender, 2022, 2024; Lindemann, 2025) is important but orthogonal to the point here.

for autonomous vehicles (*Ibid.*).

It's not always the case that data collection happens without payment, but even when data work is paid, it is typically for extremely low wages, under precarious and gigified conditions, with rampant wage theft (Fort et al., 2011; Fuentes, 2024; Okinyi, 2024; Hao, 2025). The platforms that allow for this kind of labor are presented as APIs (application programming interfaces) to people, and the low cost and ease of access to people's labor as extremely beneficial. For example, Little et al. (2010, p.57) write of Amazon's platform, "Mechanical Turk (MTurk) provides an on-demand source of human computation. This provides a tremendous opportunity to explore algorithms which incorporate human computation as a function call."

Denton et al. (2021) present a history of ImageNet (Deng et al., 2009) and examine in detail how Fei-Fei Li and her collaborators on that project leveraged labor through Amazon's Mechanical Turk and conceptualized the affordances of the platform as a way to make feasible the scaling of the collection of 'gold-standard' (i.e. provided by people) data labels. The purpose of the drive for scale was to enable a big-data approach to machine learning, as has indeed taken off. Denton et al. analyze the texts produced by the ImageNet authors about their work and show that not only was the approach about getting access to human labor inexpensively enough to enable massive scale, but also that it decontextualized and thus dehumanized the very labor that was valued for being human: "[Workers] are utilized as a generic human intelligence resource capable of executing the requested tasks of labeling images on the AMT platform. This is premised on the idea that all humans have the innate capacity to recognize images in the same way—an approach to vision that erases lived experience from the formation of meaning" (Denton et al., 2021, p.8).

The human work hidden behind the veneer of 'AI' comprises every part of the 'AI' pipeline. It is in the original media (text, image, video) appropriated and the labels cheaply commissioned to create the original training data. It is in further steps of system refinement, for example reinforcement learning from human feedback (RLHF; Ouyang et al. 2022), where data workers or end users provide evaluations of system output that are used to further set weights within large models. It is in the hidden work of people monitoring or flat-out running purportedly fully automatic systems ("fauxtomation"; Taylor 2018) including self-driving cars in the US remotely operated as needed by workers in Mexico (Kolodny, 2023), AmazonGo's "just walk out" payment system being handled by workers in India (Bitter, 2024), or supposed AI software

engineering being provided by software engineers in India (TOI Tech Desk, 2025). The work doesn't have to be hidden to be devalued either: it is also in the replacement of stable, more fulfilling jobs with other lower-paid and more precarious work fixing up the output of so-called AI systems that were deployed as ostensible replacements for the work of people. For example, despite automation being used in translation for many years, when the hype wave around ChatGPT came through, translators found their work drying up and being replaced (in part) by offers to post-edit machine translation (Merchant, 2025), work which pays less but is not correspondingly easier than translating from scratch.

As user-facing systems and services, 'AI' products are sold as cost-saving, efficient, convenient automation. As noted, however, that presentation of automation hides enormous amounts of devalued human labor. At the same time, there is enormous wealth being accumulated in the brokering of that data. For example, in late 2025, The Verge reported that Mercor, a start-up that connects labs and companies seeking specialized annotation types with gig workers with relevant expertise (including doctors, lawyers, physicists and software engineers), had reached a valuation of $10 billion, with $500 million annualized revenue (Dzieza and Field, 2025). The people who sell access to other people's devalued, gigified, precarious labor can do so at a handsome profit. This function of AI is a key ingredient in how AI is used to centralized power, as discussed in §8 below.

7 AI is a way to shift accountability

In 1976, Weizenbaum warned us against taking computer output, which can look like the answer to a consequential decision, as an appropriate way to make those decisions:

> Computers can make judicial decisions, computers can make psychiatric judgments. They can flip coins in much more sophisticated ways than can the most patient human being. The point is that they ought not be given such tasks. [...] What emerges as the most elementary insight is that, since we do not now have any ways of making computers wise, we ought not now to give computers tasks that demand wisdom. (Weizenbaum, 1976, p.227)

Despite this, automation has been integrated into decision-making processes across society in the years since, often with the goals of efficiency (making decisions faster, or handling more cases with fewer staff) or impartiality (taking human subjectivity out of decision-making). The harms of these moves have been thoroughly documented, including Eubanks's (2018) work on the "digital poorhouse", automated processes that serve to deny social benefits, and Benjamin's (2019) articulation of the "new Jim Code", systems which encode systemic racism but place them behind a veneer of objectivity.

A key aspect of the move towards automation in these contexts is that it displaces or obfuscates accountability (Nissenbaum, 1996). Whereas a person or group of people making a decision have clear responsibility for both the process and the impacts of that decision, deploying a software system for decision making renders that accountability harder to trace (though no less present). Nissenbaum identifies four factors which contribute to this obfuscation: the problem of "many hands" (all of the people involved in creating, procuring and using the system), the practice of software licenses which deny liability, the acceptance of bugs as just a normal feature of software systems, and the tendency to blame computers, as the proximal source of a problem.

In a military context, Suchman (2020) traces how algorithmic warfare conflates accuracy of striking an intended target once chosen with accuracy of choice of targets, displacing from the US military "culpability [...] in increasing reliance on ever more questionable forms of stereotypic categorisation of who constitutes a legitimate target, and the expanding temporal and spatial boundaries of what comprises an imminent threat" (p.176). Salvaggio (2025) articulates how the use of AI by DOGE (the US Department of Government Efficiency) in early 2025 was not about efficiency (as claimed) but rather about deflecting accountability: "LLMs are not just text generators but pretext generators. AI is most potent as a discursive tool to justify and enact actions for which nobody wants to be accountable."

The accountability isn't always displaced onto the systems themselves. Elish (2019) builds the analogy of "moral crumple zones": Where the crumple zone on a car is meant to take the impact in an accident to protect the passengers inside, the people who are closest to the incident (e.g. safety drivers in self-driving car tests, technicians using largely automated medical equipment, pilots flying planes with extensive auto-pilot systems) are accorded responsibility for all failures involving the system, protecting the larger organizations that build and deploy the systems.

There are several aspects of how the idea of artificial intelligence has been developed and sold which further exacerbate these problems. The anthropomorphization of systems, together with the marketing pitch that they are (on the way to being) general-purpose (e.g. Eloundou et al., 2024), suggests both functionality that might substitute for human judgement and perhaps even the ability to make decisions with accountability. The massive scale of data used in system development gives as false veneer of objectivity (Bender, Gebru et al 2021), wherein the systems are understood to have more 'knowledge' than any person could ever access and a viewpoint that is not tethered to any specific person's experience, i.e. Haraway's (1988) "view from nowhere." As McQuillan (2022, 51) writes, "AI is a form of scientism. It uses the aura of science to perpetuate the idea that its abstract mathematical models provide a reliable way of knowing, and promotes a reductive definition of truth that is claimed as inherently superior to lived experience."

As a final illustrative example, consider the government of Albania which set up a chatbot dubbed Diella to provide access to citizen services (Henley, 2025). In September 2025, Albanian Prime Minister Edi Rama announced that this chatbot would be considered a cabinet minister and used to make decisions in public tenders. The stated goal is to avoid corruption in these decisions. There are two possibilities here: The first is that the process of making these decisions is one that is fully mechanical and deterministic and thus can be automated. Alternatively, and far more likely, it is a process that requires judgment. By taking the output of the automated system as the decision in this case, the people with responsibility are shirking the duty of making the judgment calls they have been elected to make, and replacing those judgment calls with at best a random lottery but more likely a system that reflects and reproduces historical biases (Birhane et al., 2022). In sum, artificial intelligence can be used to obfuscate accountability and its usefulness as a tool for doing so is supported by the ideologies around it, as expressed through the ways in which it is imagined and marketed.

8 AI is a way to centralize power

Hao et al. (2022), Tacheva and Ramasubramanian (2023) and Hao (2025) liken AI to empire, arguing that AI companies are centralizing power analogously to the European imperial powers. They observe analogies to historical empires on many levels: in the imposition of ways of understanding

the world, in the extraction of natural resources, in the exploitation of labor, and in surveillance, especially but not only of marginalized or othered populations. This section takes each of these briefly in turn.

In imposing a specific understanding of the world, the empires or would-be empires of AI build on AI as an ideology (§5 above). Benjamin (2024) identifies not only the eugenicist ideology underlying AI, but also "eugenic *infrastructures*—systems designed to sacrifice the lives and habitats of the global majority to ensure the flourishing of the oligarchic minority" (emphasis in original). Tacheva and Ramasubramanian (2023) argue that taking the logic of AI, the idea that the world, especially the social world, can be datafied and computationally processed, is grounded in modernist/colonialist viewpoints. Citing Mumford (2022) and Rouvroy et al. (2013), they write of "'algorithmic governmentality' whereby our reality becomes shaped and controlled through the statistical probabilistic logic of AI which can be traced back to Enlightenment-era scientific principles" (p.6). The assertion, often presupposed, that AI is an inevitable future that everyone must race towards, and the idea that we should ensure that all nations achieve that same end, avoiding a worsening 'digital divide' is similarly a colonialist viewpoint, with the colonial powers representing advancement, modernity and the leading edge into the future being the tech companies serving as the competing empires of AI.

Whereas the abstract idea of thinking machines doesn't necessarily entail massive exploitation of physical resources, the way it is being carried out in the 2010s and 2020s, at ever larger scales, is deeply resource intensive. This race for scale has its roots in the 1980s, when Robert Mercer, an early proponent of data-driven as opposed to knowledge-engineered approaches to speech and language processing, declared "There's no data like more data" (Withgott and Chen, 1993, p.81). Data-driven approaches in general take off within natural language processing with the establishment (in 1993) and growth of SIGDAT, the Association for Computational Linguistics' Special Interest Group on Linguistics Data & Corpus-based Approaches to Natural Language Processing.[8] In image processing, the ImageNet project (Deng et al. 2009; see §6) inspired an era of big-data approaches. An important inflection point came when Krizhevsky et al. (2012) showed how deep (large) neural networks could be used to take advantage of the scale of input data provided by ImageNet. This kicked off a race towards ever-larger models

[8] `https://sigdat.org/about`

(in terms of both input data and compute) first in image processing then in speech and language processing (Bender, Gebru et al 2021).

This race, despite being nominally about abstract objects (data sets, algorithms, models) has very real material consequences, as it has motivated and demanded a race to build more and more data centers. Colliers (2025) reports that global data center investment was $26 billion in 2023 and $57 billion in 2024 (with another $29 billion pending at the end of 2024). The construction of the data centers themselves is resource intensive, demanding rare earth minerals to create the chips as well as high fossil fuel and other chemical inputs, and their associated greenhouse gas emissions and toxic waste (Williams, 2004; Kim et al., 2024). The running of data centers is also environmentally expensive, demanding large amounts of reliably available energy and fresh water (Hao, 2025). The International Energy Association estimates that global supply of electricity to power data centers was 460 TWh in 2024 (IEA, 2025). That includes all uses of data centers, but the push to exponentially larger data centers ('hyperscale') is driven by the demand for compute coming from large models called AI (Marx, 2024).

The labor practices described in §6 above are not only exploitative but also colonial (Hao and Hernández, 2022). The platforms and subcontractors that enable tech companies to access data workers' labor tend to source that labor from among populations experiencing economic and other crises, at wages that are barely sustainable or unsustainable and under working conditions that keep workers tethered to their computers competing to catch the best paying tasks as soon as they are posted. As Hao (2022) writes, this is based on "implicit ideas that such populations don't need — or are less deserving of — livable wages and economic stability."

Empires require containment and control of the colonized populations. In the context of empires of AI, this plays out through surveillance. Technologies billed as and developed under the rubric of 'AI' require and normalize surveillance while also serving as technologies for surveillance. Couldry and Mejias (2019) elaborate a concept of data colonialism, the appropriation of human life and experience for extractive purposes. Zuboff (2019) documents how the tech industry has conceptualized digital spaces as *terra incognita*, available for claiming and draws an explicit analogy to European powers colonizing what they considered the New World. Those digital spaces include large and increasing amounts of data about people: our writings and photos that we post online, our behavior in online spaces (which links we click, which posts we like, etc.), as well as biometric information, often given

up under duress (Molnar, 2024) or otherwise without consent. Sometimes, the relinquishing of personal data is the price of entry for the use of convenient consumer goods, what Gilliard and Golumbia (2021) term "luxury surveillance". Gilliard and Golumbia note that the technologies developed in this way aren't confined to use by the people who opt in (and pay to do so), but support the development of coercive surveillance technologies, drawing for example the analogy between GPS-enabled fitness trackers and ankle bracelets imposed on parolees.

More generally, technologies developed under the moniker of 'artificial intelligence' are frequently, if not overwhelmingly, developed and used for the purpose of surveillance by governments and other institutions. Buolamwini (2023) and the Algorithmic Justice League[9] document the ways in which facial recognition technologies are used for surveillance, with high (even life-and-death) stakes both when they are accurate and when they fail. Kalluri et al. (2025) survey papers from 1990-2021 in image and video processing and patents related to them and find that 88% involve handling data about humans in image/video data (including human bodies, human body parts, and human spaces). Furthermore they find that in the 1990s, papers used in surveillance-enabling patents were roughly equal in number of papers used only in non-surveillance enabling patents, but in the 2010s, the former increased five-fold while the latter held almost steady.

In building up and building on logics of empire, artificial intelligence serves as a means of centralizing power. Power thus centralized accrues in the first instance to the companies that control the data and computational infrastructure. However, it also figures in the struggle between nation states, for example, in the appeals to an AI arms race with China in documents such as US Senator Chuck Schumer et al's "Roadmap for artificial intelligence policy in the United States Senate" (Schumer et al., 2024) or the Trump administration's AI Action Plan (Kratsios et al., 2025). In reproducing and reinscribing the hegemonic worldview predominant in the training data (Bender, Gebru et al 2021; Birhane et al. 2022) and colonial ideologies more generally, it also accrues to anyone who stands to benefit from those power structures.

[9] `https://www.ajl.org/`

9 Conclusion

The notion of artificial intelligence is frequently sold as present or near-future and inevitable technology. In fact there is no coherent set of technologies that can serve as the denotation of the phrase, nor do any of the technologies so marketed rise to the fantastical but ill-defined claims of 'AI' is or soon will be. Nonetheless, the idea of artificial intelligence has been extremely impactful in the world. In order to better understand and deal with those impacts, it is helpful to look at artificial intelligence through the varied lenses of how the idea operates in the world: as the name of a research field, as one approach to cognitive science, as a parlor trick, as a an ideology, as a way to hide and devalue human labor, as a way to shift and/or obfuscate accountability, and as a means to centralize power.

10 Further Reading

References

Bender, E. M. and A. Hanna (2025). *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. New York: HarperCollins.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Polity Press.

Buolamwini, J. (2023). *Unmasking AI: My Mission to Protect What is Human in a World of Machines*. New York: Penguin Random House.
Data Worker's Inquiry

Fort, K., G. Adda, and K. B. Cohen (2011, June). Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics 37*(2), 413–420.

Gebru, T. and E. P. Torres (2024, Apr.). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday 29*(4).

Hao, K. (2025). *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. New York: Penguin Random House.

Marx, P. (2024). Data vampires. Four part podcast series, from Tech Won't Save Us, available at https://techwontsave.us/, episodes 241, 243, 245, and 247.

Molnar, P. (2024). *The Walls Have Eyes: Surviving Migration in the Age of Artificial Intelligence*. The New Press.

Paullada, A., I. D. Raji, E. M. Bender, E. Denton, and A. Hanna (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns 2*.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco CA: Freeman.

References

Abdurahman, J. K. (2022). Birthing predictions of premature death. *Logic(s)*.

Agüera y Arcas, B., M. Mitchell, and A. Todorov (2023). Physiognomy in the age of AI. In *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, pp. 208–236. Oxford Academic.

Alkhatib, A. (2024). Defining AI. Blog post, accessed 6 Sept 2025, https://ali-alkhatib.com/blog/defining-ai.

ALPAC (1966). Language and machines; computers in translation and linguistics. Technical report, National Academy of Sciences, Publication 1416, Washington DC.

Altman, S. s. (2022). "i am a stochastic parrot, and so r u.". Twitter, December 4, 2022, https://x.com/sama/status/1599471830255177728, accessed September 6, 2025.

Amodei, R. (2024). Machines of loving grace: How AI could transform the world for the better. Blog post, accessed 31 August 2025, https://www.darioamodei.com/essay/machines-of-loving-grace.

Baria, A. T. and K. Cross (2021). The brain is a computer is a brain: Neuroscience's internal debate and the social significance of the computational metaphor. *arXiv*.

Beer, S. (2002, 03). What is cybernetics? *Kybernetes 31*(2), 209–219.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, New York, NY, USA, pp. 610–623. Association for Computing Machinery.

Bender, E. M. and A. Hanna (2025). *The AI Con: How to Fight Big Tech's Hype and Create the Future We Want*. New York: HarperCollins.

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge, UK: Polity Press.

Benjamin, R. (2024). The new artificial intelligentsia. *Los Angeles Review of Books*.

Birhane, A., P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao (2022). The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, New York, NY, USA, pp. 173184. Association for Computing Machinery.

Bitter, A. (2024). Amazon's just walk out technology relies on hundreds of workers in India watching you shop. *Business Insider*.

Blili-Hamelin, B., C. Graziul, L. Hancox-Li, H. Hazan, E.-M. El-Mhamdi, A. Ghosh, K. A. Heller, J. Metcalf, F. Murai, E. Salvaggio, A. J. Smart, T. Snider, M. Tighanimine, T. Ringer, M. Mitchell, and S. Dori-Hacohen (2025). Position: Stop treating 'AGI' as the north-star goal of AI research. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Bowers, J. S., G. Malhotra, F. Adolfi, M. Dujmovi, M. L. Montero, V. Biscione, G. Puebla, J. H. Hummel, and R. F. Heaton (2023). On the importance of severely testing deep learning models of cognition. *Cognitive Systems Research 82*, 101158.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics 19*(2), 263–311.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Browne, R. (2025). AI that can match humans at any task will be here in five to 10 years, Google DeepMind CEO says. *CNBC*.

Buolamwini, J. (2023). *Unmasking AI: My Mission to Protect What is Human in a World of Machines*. New York: Penguin Random House.

Chayka, K. (2023). Is A.I. art stealing from artists? *The New Yorker*.

Chiang, W.-L., L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. I. Jordan, J. E. Gonzalez, and I. Stoica (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Cognilytica (2020). AI data engineering lifecycle checklist: Following steps for AI project success. White paper, accessed Sept 8, 2025, available at `https://www.cloudera.com/content/dam/www/marketing/resources/whitepapers/ai-data-lifecycle-checklist-cloudera-whitepaper.pdf?daqp=true`.

Colliers (2025). 2025 data center marketplace: Balancing unprecedented opportunity with strategic risk. Available at `https://www.colliers.com/download-article?itemId=55e5a5a6-c48f-4ac3-ac9d-cf7d0f5325e7`.

Couldry, N. and U. A. Mejias (2019). *The Costs of Connection: How Data is Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press.

Davis, K. H., R. Biddulph, and S. Balashek (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America 24*(6), 637–642.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE.

Denton, R., A. Hanna, R. Amironesei, A. Smart, and H. Nicole (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society 8*(2), 1–14.

Dzieza, J. and H. Field (2025). Feeding the machine. *The Verge*. December 15, 2025.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society 5*, 40–60.

Eloundou, T., S. Manning, P. Mishkin, and D. Rock (2024). GPTs are GPTs: Labor market impact potential of LLMs. *Science 384*(6702), 1306–1308.

Eubanks, V. (2018). *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Feigenbaum, E. A. (1963). Artificial intelligence research. *IEEE Transactions of the Professional Technical Group on Information Theory 1T-9*(4).

Fort, K., G. Adda, and K. B. Cohen (2011, June). Last words: Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics 37*(2), 413–420.

Fuentes, O. V. (2024). Life of a Latin American data worker. In M. Miceli, A. Dinika, K. Kauffman, C. S. Wagner, and L. Sachenbacher (Eds.), *The Data Workers' Inquiry,* `https://data-workers.org/oskarina`.

Gebru, T. and E. P. Torres (2024, Apr.). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday 29*(4).

Gilliard, C. and D. Golumbia (2021). Luxury surveillance: People pay a premium for tracking technologies that get imposed unwillingly on others. *Real Life Mag*.

Goldberg, D., D. Nichols, B. M. Oki, and D. Terry (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM 35*(12), 61–70.

Gould, S. J. (1981). *Mismeasure of Man*. WW Norton & company.

Guest, O. and A. E. Martin (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science 16*(4), 789–802. PMID: 33482070.

Hao, K. (2022). Artificial intelligence is creating a new colonial world order. *MIT Technology Review*.

Hao, K. (2025). *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. New York: Penguin Random House.

Hao, K. and A. P. Hernández (2022). How the AI industry profits from catastrophe. *MIT Technology Review*.

Hao, K., H. Swart, A. P. Hernández, and N. Freischlad (2022). AI colonialism. *MIT Technology Review*.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies 14*(3), 575–599.

Hassenfeld, N. (2023). Even the scientists who build AI can't tell you how it works. *Unexplainable Podcast, Vox.com*.

He, P., X. Liu, J. Gao, and W. Chen (2021). Microsoft DeBERTa surpasses human performance on the superglue benchmark.

Henley, J. (2025). Albania puts AI-created 'minister' in charge of public procurement. *The Guardian*.

Hu, K. and N. Nishant (2025). US AI startups see funding surge while more VC funds struggle to raise, data shows. *Reuters*. Juy 15, 2025.

Hullman, J., S. Kapoor, P. Nanayakkara, A. Gelman, and A. Narayanan (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, New York, NY, USA, pp. 335348. Association for Computing Machinery.

IEA (2025). Energy and ai. Available at `https://www.iea.org/reports/energy-and-ai`, Licence: CC BY 4.0.

Jamali, L. and L. McMahon (2025). OpenAI claims GPT-5 model boosts ChatGPT to 'phd level'. *BBC*.

Jia, S. and A. Nagaraj (2025). The impact of pirated data on large language model performance: A study on Books3.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 596*(7873), 583–589.

Kalluri, P. R., W. Agnew, M. Cheng, K. Owens, L. Soldaini, and A. Birhane (2025). Computer-vision research powers surveillance technology. *Nature*, 1–7.

Kim, G. C., A. Rothschild, C. DiSalvo, and B. DiSalvo (2024, Oct.). What's your stake in sustainability of AI?: An informed insider's guide. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 7*(1), 738–750.

Kolodny, L. (2023). Cruise confirms robotaxis rely on human assistance every four to five miles. *CNBC*.

Kratsios, M. J., D. O. Sacks, and M. A. Rubio (2025). Winning the race: America's AI action plan. Report from the Trump White House, Available at `https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf`.

Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*.

Levinson, S. C. (2025). *The Interaction Engine: Language in Social Life and Human Evolution*. Cambridge University Press.

Lighthill, J. (1973). Artificial intelligence: A general survey. *Science Research Council (SRC), Government Report*.

Lindemann, N. F. (2025). Chatbots, search engines, and the sealing of knowledges. *AI & Society 40*, 5063–5076.

Little, G., L. B. Chilton, M. Goldman, and R. C. Miller (2010). TurKit: Human computation algorithms on Mechanical Turk. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, New York, NY, USA, pp. 57–66. Association for Computing Machinery.

LMArena (2025). About us. Web page, accessed September 11, 2025. `https://lmarena.ai/about`.

Marshall, J. C. (1977). Minds, machines and metaphors. *Social Studies of Science 7*(4), 475–488.

Marx, P. (2024). Data vampires. Four part podcast series, from Tech Won't Save Us, available at `https://techwontsave.us/`, episodes 241, 243, 245, and 247.

Maslej, N., L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, T. Walsh, A. Hamrah, L. Santarlasci, and J. B. Lotufo (2025). The AI index 2025 annual report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025. Available at `https://hai.stanford.edu/ai-index/2025-ai-index-report`.

McCarthy, J. (1990). Chess as the drosophila of AI. In T. A. Marsland and J. Schaeffer (Eds.), *Computers, Chess, and Cognition*, New York, NY, pp. 227–237. Springer New York.

McCarthy, J., M. Minsky, N. Rochester, and C. Shannon (1955). A proposal for the Dartmouth summer research project on artificial intelligence. Research project proposal, pdf available at `http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf`, accessed 1 September 2025.

McQuillan, D. (2022). *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. Bristol, UK: Bristol University Press.

Merchant, B. (2025). AI killed my job: Translators. *Blood in the Machine* newsletter, `https://www.bloodinthemachine.com/p/ai-killed-my-job-translators`, accessed September 9, 2025.

Molnar, P. (2024). *The Walls Have Eyes: Surviving Migration in the Age of Artificial Intelligence*. The New Press.

Mumford, D. (2022). Data colonialism: Compelling and useful, but whither epistemes? *Information, Communication & Society 25*(10), 1511–1516.

Narayanan, A. and S. Kapoor (2024). *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton University Press.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics 2*(1), 25–42.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Okinyi, M. (2024). Impact of Remotasks closure on Kenyan workers. In M. Miceli, A. Dinika, K. Kauffman, C. S. Wagner, and L. Sachenbacher (Eds.), *The Data Workers' Inquiry,* `https://data-workers.org/mophat`.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), *Advances in Neural Information Processing Systems*, Volume 35, pp. 27730–27744. Curran Associates, Inc.

Paullada, A., I. D. Raji, E. M. Bender, E. Denton, and A. Hanna (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns 2*.

Pettis, B. T. (2023). reCAPTCHA challenges and the production of the ideal web user. *Convergence 29*(4), 886–900.

Raji, D. (2020). How our data encodes systematic racism. *MIT Technology Review*.

Raji, I. D., E. M. Bender, A. Paullada, E. Denton, and A. Hanna (2021). AI and the everything in the whole wide world benchmark. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.

Raji, I. D., I. E. Kumar, A. Horowitz, and A. Selbst (2022). The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, New York, NY, USA, pp. 959972. Association for Computing Machinery.

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In A. Ortony (Ed.), *Metaphor and Thought*, pp. 164–201. Cambridge University Press.

Rouvroy, A., T. Berns, and L. Carey-Libbrecht (2013). Algorithmic governmentality and prospects of emancipation. *Réseaux 177*(1), 163–196.

Salvaggio, E. (2025). Musk, AI, and the weaponization of 'administrative error'. *Tech Policy Press*.

Schumer, C., M. Rounds, M. Heinrich, and T. Young (2024). Driving U.S. innovation in artificial intelligence: A roadmap for artificial intelligence policy in the united states senate. Report of the Bipartisan Senate AI Working Group, available at `https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf`.

Scott, J. C. (1998). *Seeing Like A State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven, CT: Yale University Press.

Shah, C. and E. M. Bender (2022). Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, New York, NY, USA, pp. 221–232. Association for Computing Machinery.

Shah, C. and E. M. Bender (2024). Envisioning information access systems: What makes for good tools and a healthy web? *ACM Trans. Web 18*(3).

Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis (2017). Mastering the game of Go without human knowledge. *Nature 550*(7676), 354–359.

Smith, C. (1993). The use and abuse of metaphors in the history of brain science. *Journal of the History of the Neurosciences 2*(4), 283–301. PMID: 11618462.

Stark, L. and J. Hutson (2022). Physiognomic artificial intelligence. *Fordham Intellectual Property, Media and Entertainment Law Journal 32*(4).

Suchman, L. (2020). Algorithmic warfare and the reinvention of accuracy. *Critical Studies on Security 8*(2), 175–187.

Sweeney, L. (May 1, 2013). Discrimination in online ad delivery. *Communications of the ACM 56*(5), 44–54.

Tacheva, J. and S. Ramasubramanian (2023). AI empire: Unraveling the interlocking systems of oppression in generative AI's global order. *Big Data & Society 10*(2), 20539517231219241.

Taigman, Y., M. Yang, M. Ranzato, and L. Wolf (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708.

Tarnoff, B. (2023). Computer scientist Joseph Weizenbaum was there at the dawn of artificial intelligence — But he was also adamant that we must never confuse computers with humans. *The Guardian*. June 25, 2023.

Taylor, A. (2018). The automation charade. *Logic(s) Magazine*.

TOI Tech Desk (2025). How this Microsoft-backed billion-dollar London startup made 700 engineers sitting in India pose as AI tool. *Times of India*.

Toole, B. A. (1998). *Ada, the Enchantress of Numbers: Prophet of the Computer Age, a Pathway to the 21st Century*. Sausalito, CA: Critical Connection.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind: A Quarterly Review of Psychology and Philsophy 59*(236), 433–460.

van Rooij, I. and G. Baggio (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science 16*(4), 682–697. PMID: 33404356.

van Rooij, I., M. Blokpoel, J. Kwisthout, and T. Wareham (2019). *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis*. Cambridge University Press.

van Rooij, I., O. Guest, F. Adolfi, R. de Haan, A. Kolokolova, and P. Rich (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior 7*(4), 616–636.

von Ahn, L., M. Blum, N. J. Hopper, and J. Langford (2003). CAPTCHA: Using hard AI problems for security. In E. Biham (Ed.), *Advances in Cryptology — EUROCRYPT 2003*, Berlin, Heidelberg, pp. 294–311. Springer Berlin Heidelberg.

von Ahn, L. and L. Dabbish (2008, August). Designing games with a purpose. *Commun. ACM 51*(8), 58–67.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pp. 3266–3280.

Weizenbaum, J. (1966). Eliza — a computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*(1), 36–45.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco CA: Freeman.

Williams, E. D. (2004). Environmental impacts of microchip manufacture. *Thin Solid Films 461*(1), 2–6.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology 3*(1), 1–191.

Withgott, M. M. and F. R. Chen (1993). *Computational Models of American Speech*. Number 32. Center for the Study of Language (CSLI).

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs Books.