Position: Stop Treating ‘AGI’ as the North-star Goal of AI Research

Borhane Blili-Hamelin [* 1] **Christopher Graziul** [* 2] **Leif Hancox-Li** [† 3] **Hananel Hazan** [† 4]
El-Mahdi El-Mhamdi [‡ 5 6] **Avijit Ghosh** [‡ 7 8] **Katherine Heller** [‡ 9] **Jacob Metcalf** [‡ 10] **Fabricio Murai** [‡ 11]
Eryk Salvaggio [‡ 12] **Andrew Smart** [‡ 9] **Todd Snider** [‡ 13] **Mariame Tighanimine** [‡ 14] **Talia Ringer** [§ 15]
Margaret Mitchell [§ 7] **Shiri Dori-Hacohen** [§ 8]

Abstract

The AI research community plays a vital role in shaping the scientific, engineering, and societal goals of AI research. In this position paper, we argue that focusing on the highly contested topic of ‘artificial general intelligence’ (‘AGI’) undermines our ability to choose effective goals. We identify six key traps—obstacles to productive goal setting—that are aggravated by AGI discourse: Illusion of Consensus, Supercharging Bad Science, Presuming Value-Neutrality, Goal Lottery, Generality Debt, and Normalized Exclusion. To avoid these traps, we argue that the AI research community needs to (1) prioritize **specificity** in scientific, engineering, and societal goals, (2) center **pluralism** about multiple worthwhile approaches to multiple valuable goals, and (3) foster innovation through greater **inclusion** of disciplines and communities. Therefore, the AI research community needs to **stop treating “AGI” as the north-star goal of AI research**.

[*]Equal contribution & Lead contributors [†]Top contributors [‡]Contributors (alphabetical order) [§]Senior contributors [1]AI Risk and Vulnerability Alliance, NY, USA [2]University of Chicago, IL, USA [3]Vijil, CA, USA [4]Tufts University, MA, USA [5]Calicarpa, Switzerland [6]Ecole Polytechnique, France [7]Hugging Face, NY, USA [8]University of Connecticut, CT, USA [9]Google, CA, USA [10]Data & Society Research Institute, NY, USA [11]Worcester Polytechnic Institute, MA, USA [12]Rochester Institute of Technology, NY, USA [13]Eberhard Karls Universität Tübingen, Tübingen, Germany [14]Conservatoire national des arts et métiers, Lise-CNRS, France [15]University of Illinois Urbana-Champaign, IL, USA. Correspondence to: Borhane Blili-Hamelin <borhane@avidml.org>, Christopher Graziul <graziul@uchicago.edu>, Talia Ringer <tringer@illinois.edu>, Margaret Mitchell <meg@huggingface.co>, Shiri Dori-Hacohen <shiridh@uconn.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

How can we ensure that AI research goals serve scientific, engineering, and societal needs? What constitutes good science in AI research? Who gets to shape AI research goals? What makes a research goal legitimate or worthwhile? In this position paper, we argue that a widespread emphasis on AGI threatens to undermine the ability of researchers to provide well-motivated answers to these questions.

Recent advances in large language models (LLMs) have sparked interest in “achieving human-level ‘intelligence”’ as a “north-star goal” of the AI field (McCarthy et al., 1955; Morris et al., 2024). This goal is often referred to as “artificial general intelligence” (“AGI”) (Chollet, 2024a; Tibebu, 2025). Yet rather than helping the field converge around shared goals, AGI discourse has mired it in controversies. Researchers diverge on what AGI is and assumptions about goals and risks (Summerfield, 2023; Morris et al., 2024; Blili-Hamelin et al., 2024). Researchers further contest the motivations, incentives, values, and scientific standing of claims about AGI (Gebru & Torres, 2024; Mitchell, 2024; Ahmed et al., 2024; Altmeyer et al., 2024). Finally, the building blocks of AGI as a concept—intelligence and generality—are contested in their own right (Gould, 1981; Anderson, 2002; Hernández-Orallo & Seán Ó hÉigeartaigh, 2018; Cave, 2020; Raji et al., 2021; Alexandrova & Fabian, 2022; Blili-Hamelin & Hancox-Li, 2023; Hao, 2023; Guest & Martin, 2024; Paolo et al., 2024; Mueller, 2024).

Building on prior work on the ambiguity between exploratory and confirmatory research in ML (Herrmann et al., 2024), unscientific performance claims (Altmeyer et al., 2024), SOTA-chasing (Raji et al., 2021; Church & Kordoni, 2022), homogenization of research approaches (Kleinberg & Raghavan, 2021; Fishman & Hancox-Li, 2022; Bommasani et al., 2022), the values embedded in ML research (Birhane et al., 2022b), and more, our account identifies key obstacles to productive goal setting in AI research—**traps**.[1] We provide them here as a diagnosis of problems in goal-setting

[1]Our terminology parallels Selbst et al. (2019) on fairness.

that we believe are normatively worth addressing, but that the AGI narrative makes difficult to overcome. To avoid these traps, **we posit that communities should stop treating AGI as the north-star goal[2] of AI research**.

An overarching theme in our discussion is the research community's *unique responsibility to help distinguish hype from reality*. The outputs of AI research are deployed as real-world products at a staggering pace, in proliferating contexts, affecting billions of people. This warrants urgent work on trusted, evidence-based answers to questions about the scientific, engineering, and societal merits of AI tools. As argued by the U.N.'s AI Advisory Body, there is "an overwhelming amount of information… making it difficult to decipher hype from reality. This can fuel confusion, forestall common understanding and advantage major AI companies at the expense of policymakers, civil society and the public" (United Nations, 2024). Our position paper addresses this theme: Each of the six traps in our account is an obstacle to distinguishing hype from reality.

A secondary theme in our discussion is the relationship between people and technology. Ultimately, we argue that instead of a single north-star goal, the AI community needs to pursue *multiple specific* scientific, engineering, and societal goals. If building consensus around an alternative unifying goal proves useful, we propose the goal of *supporting and benefiting human beings*.

In the next section, we examine six 'traps' in AI research—obstacles to productive goal setting (§2). We argue that AGI discourse reinforces and amplifies each problem. Subsequently, we provide three recommendations for avoiding these traps (§3): specificity of goals; pluralism of goals and approaches; and more inclusive goal setting. We conclude by offering a rebuttal against an *alternative view* (§4): that AGI should remain the north-star goal of the field.

Because the contested nature of AGI is a central theme in the present paper, we avoid providing our own definition for the term. Instead, we provide example definitions throughout the present discussion, as well as a table of illustrative definitions in Appendix A.[3] We detail how AGI currently serves as a north-star goal in Appendix B.

[2]Sailors who navigate by the astronomical North Star use it to orient their travels toward a desired destination on Earth. With AGI, some researchers are using AGI as a "guiding star" to orient their AI research "travels" towards. Other researchers, however, are actually hoping and working towards the goal of "arriving" at AGI. Our paper argues against both of these approaches.

[3]Arguably, many of the concerns we raise about AGI apply to other terms used to refer to future forms of AI, such as "powerful AI" (Amodei, 2024) and "transformative AI" (Gruetzemacher & Whittlestone, 2022). Ultimately, as we revisit in §4, our account can be viewed as critically interrogating north-star goals more generally.

2. Traps

We examine six key *traps* that hinder the research community's ability to set worthwhile goals. We argue that each is aggravated by AGI narratives.

The problems we discuss are highly interrelated. For instance, SOTA-chasing, discussed in relationship to the role of misaligned incentives in shaping goal setting (§2.4), also has implications for bad science (§2.2). Similarly, the problem of the lack of consensus about AGI (§2.1) is a theme that recurs throughout. We do not intend the traps to be mutually exclusive. Rather, our goal for each trap is to provide distinct and useful insights for mitigating failure modes in productive goal setting.

2.1. Illusion of Consensus

Using shared term(s) in a way that gives a false impression of consensus about goals, despite goals being contested

The popular use of the term "AGI" (Grossman, 2023; IBM, 2023; Holland, 2025) creates a sense of familiarity, giving the illusion that there is a shared understanding on what AGI is, and broad agreement on research goals in AGI development. However, there are vastly different opinions on what the term AGI refers to, what an AGI research agenda looks like, and what the goals in AGI development are. Left unchecked, this illusion obstructs explicit engagement on what the goals of AI research are and should be.

From popular discourse to research papers to corporate marketing materials, the vast majority of references to AGI fall into this trap when they uncritically cite claims about so-called AGI. For examples of uncritical media claims, see Grossman (2023) and IBM (2023); see Altmeyer et al. (2024) for examples of overhyped research.[4] Summerfield (2023) summarizes the issue: "AI researchers hope to discover how to build AGI. The problem is that nobody really knows exactly what an AGI would look like." Mueller (2024) calls AGI "a meaningless concept, an emperor with no clothes." Blili-Hamelin et al. (2024) identify multiple types of disagreement among definitions of AGI or human-level AI. The contested nature of AGI as a goal is even more acute in critiques of AGI concepts (e.g., Altmeyer et al., 2024; Mueller, 2024; Van Rooij et al., 2024).

Beyond AGI, AI research is rife with topics that involve disagreement about goals, values, and concepts. For example, Mulligan et al. (2016) argue that *privacy* should be understood as an "essentially contested concept." They argue lack of agreement about the meaning and significance of privacy is not merely a matter of confusion—rather, dis-

[4]Some researchers who advocate for AGI as a goal have avoided the Illusion of Consensus trap (§2.1); e.g., Morris et al. (2024) explicitly call for investigating disagreements about goals, predictions, and risks that underpin prominent accounts of AGI.

agreement and contestation are desirable features that enable privacy to adapt to changing technical and social contexts. Similarly, there is now widespread acceptance *fairness* should be understood as a contested topic, not only admitting incompatible mathematical formalizations but also incompatible values, worldviews, and theoretical assumptions (Friedler et al., 2021; Jacobs & Wallach, 2021).

In suggesting this trap, we do not presume that contested topics are inherently problematic. Rather, we argue that when dealing with the important question of the goals of AI research, the significant disagreements that surround AGI should be embraced as signals of conflicting values.

2.2. Supercharging Bad Science

Worsening current problems with bad science in AI due to poorly defined concepts and experimental procedures

Research that produces reliable empirical knowledge about AI is vital to public interest decisions about AI's potential for societal and environmental benefit and harm. Yet, many experts have noted a pervasive lack of scientific grounding in AI research (Hullman et al., 2022; Raji et al., 2022; Sloane et al., 2022; Suchman, 2023; Guest & Martin, 2024; Narayanan & Kapoor, 2024; United Nations, 2024; Van Rooij et al., 2024; Widder & Hicks, 2024). We argue that vagueness in AGI discourse exacerbates existing problems with the scientific validity of AI research.

Problem 1: Underspecification and external validity. One problem with the pursuit of AGI as a concrete goal is **underspecification** (D'Amour et al., 2022), where *lack of specificity in goals or concepts leads to cascading epistemic problems*, including irrefutability, lack of external validity, flawed experimental design, and flawed evaluation. These common problems in AI research are worsened in the AGI context by the lack of scientifically grounded definitions of AGI (§2.1).

Underspecification of learning goals also undermines *external validity*—the question of whether a measurement corresponds to the real-world phenomenon it's supposed to capture. A good example is the debate about whether "language understanding" benchmarks actually measure language understanding (Jacobs & Wallach, 2021; Liao et al., 2021).

External validity is also relevant when researchers equate human faculties with model proxies (Hullman et al., 2022), such as claiming that a model "capable of linking specific objects with more general visual context" is evidence of "imagination" (Fei et al., 2022). This rhetorical move is enabled by using colloquial terms like "imagination" without considering whether it corresponds to the human faculty. Altmeyer et al. (2024) likewise critiques Gurnee & Tegmark (2024) for inflated claims enabled by the vagueness of the term "world model". Underspecified goals trickle down into many areas of experimental design, such as learning pipelines, evaluation metrics, tasks, representations, and methods.

External validity is also undermined when researchers claim to measure concepts from other fields, like intelligence. The fields of psychology, neuroscience, and cognitive science have studied human intelligence for generations, yet even they lack consensus on what "intelligence" is (Gopnik, 2019; Hao, 2023). Conversely, AI research is no longer concerned with modeling human cognition (Guest & Martin, 2024; Van Rooij et al., 2024). Instead, AI developers define "intelligence" on their own terms, privileging definitions convenient for benchmarking or selling products (§2.4), while benefiting from historically positive connotations of the term "intelligence".

Problem 2: Ambiguity between science and engineering. Another problem with the pursuit of AGI is *confusion between science and engineering* (Agre, 2014; Hutchinson et al., 2022; Altmeyer et al., 2024). As Hullman et al. (2022) point out, a "typical supervised ML paper" (e.g., one that reports accuracy metrics on a benchmark) is often just an "engineering artifact", a tool attached to performance claims that cannot be refuted because of replication challenges. Altmeyer et al. (2024) argue that this ambiguity between science and engineering means rigorous hypothesis testing with "specific conditions and considering effect sizes" is often omitted, with results often presented as "engineering achievements" without specifying *precisely* what is being tested, relevant hypotheses, and what effect sizes would constitute substantial findings.

This ambiguity invites experimenter and confirmation biases, since researchers are incentivized to "pay little or no attention to competing hypotheses or explanations" or "[fail] to articulate a sufficiently strong null hypothesis," (Altmeyer et al., 2024). Confusion between science and engineering also manifests when it is unclear if a study is pursuing scientific goals—of explanation, hypothesis confirmation, etc.—or goals of specific engineering applications—e.g., a proof-of-concept (Hutchinson et al., 2022). This exacerbates questions about external validity: without clear and specific experimental goals, it is easier to provide post-hoc interpretations of experiments that "support" a wide variety of goals (§2.4).

Problem 3: Ambiguity between confirmatory and exploratory research. The ambiguity between engineering and scientific methodology is related to another problem: *confusion between confirmatory and exploratory research* (Bouthillier et al., 2019; Herrmann et al., 2024). Herrmann et al. (2024) state that confirmatory research "aims to test preexisting hypotheses to confirm or refute existing theories [while] exploratory research is an openended approach that aims to gain insight and understanding in a new or unexplored area." They go on to argue that "most current empirical machine learning research is fashioned as confirmatory research while it should rather be considered exploratory" and that experiments are "set up to *confirm* the (implicit) hypothesis that the proposed method constitutes an improvement" (emphasis theirs). By implicitly conflating exploratory analysis with confirmatory research, "exploratory findings have a slippery way of 'transforming' into planned findings as the research process progresses" (Calin-Jageman & Cumming, 2019). Using the vague and contested concept of AGI to frame confirmatory claims worsens this problem, as it makes it harder to figure out *what* is being claimed.

2.3. Presuming Value-Neutrality

Framing goals as purely technical or scientific, when they are in fact laden with political, social, or ethical values

Presuming Value-Neutrality occurs when technical or scientific goals become disconnected from their **value-laden** assumptions: aspects of AI research that are—and should be—informed by political, social, and ethical considerations. The AI research community has recently begun examining these value-laden assumptions (Shilton, 2018; Broussard et al., 2019; Abebe et al., 2020; Blodgett et al., 2020; Costanza-Chock, 2020; Denton et al., 2020; 2021; Dotan & Milli, 2020; Birhane & Guest, 2021; Green, 2021; Scheuerman et al., 2021; Viljoen, 2021; Birhane et al., 2022b; Bommasani, 2023; Fishman & Hancox-Li, 2022; Hutchinson et al., 2022; Mathur et al., 2022; Blili-Hamelin & Hancox-Li, 2023; Blili-Hamelin et al., 2024; Zhao et al., 2024).

When efforts to define AGI and related concepts do not explicitly examine the societal goals and values embedded in their definitions, they fall into the Presuming Value-Neutrality trap. Examples include proposals for "universal intelligence" (Legg & Hutter, 2007; Hernández-Orallo et al., 2014).

The pursuit of value-neutral approaches echoes debates about psychometric views of human intelligence. Intelligence, like "health," and "well-being," inherently carries normative assumptions about which behaviors or abilities are desirable (Anderson, 2002; Alexandrova & Fabian, 2022). Researchers fall into the Presuming Value-Neutrality trap by sidestepping these value-laden dimensions (Anderson, 2002; Cave, 2020; Blili-Hamelin & Hancox-Li, 2023). Warne & Burningham (2019) exemplify this by advocating for purely statistical definitions of intelligence, precisely because cultural definitions vary.

Value-laden assumptions within concepts like AGI drive legitimate disagreement about their meaning, reflecting divergent societal goals (Blili-Hamelin et al., 2024). This makes consensus on AGI challenging, as it requires alignment on political, social, and ethical priorities. Similar disagreements affect related concepts like AI (Cave, 2020; Blili-Hamelin & Hancox-Li, 2023), "human-level AI", "superintelligence", and "strong AI", reinforcing the Illusion of Consensus trap (§2.1).

2.4. Goal Lottery

Adopting goals which are not adequately justified by scientific, engineering, or social merit, but instead on the basis of incentives, circumstances, or luck

Researchers have studied the role of socioeconomic factors, trends, and circumstantial factors in shaping AI research. For instance, Hooker (2021) has argued that a form of hardware lottery—the greater availability of hardware with strengths in parallel processing—was key to the resurgence of deep learning in the 2010s.[5] Similarly, researchers have examined the role of incentives, socioeconomic factors, and hype cycles in AI research (Raji et al., 2022; Delgado et al., 2023; Sartori & Bocca, 2023; Widder & Nafus, 2023; Gebru & Torres, 2024; Hicks et al., 2024; Narayanan & Kapoor, 2024; Wang et al., 2024). With this trap, we focus on cases where lotteries (luck) or incentives drive the adoption of unjustified goals—goals that are inadequately supported by scientific, engineering, or societal merit.

Consider AGI definitions centered on economic value, like OpenAI's emphasis on "outperform[ing] humans at most economically valuable work" (OpenAI, 2018). The primacy of economic value for setting AI research goals is contentious from both engineering and societal perspectives. Such definitions create misalignment between incentives and justifications by reducing complex societal, engineering, and scientific considerations to purely economic metrics.

Another example is benchmark SOTA-chasing—pursuing top scores on popular benchmarks (Bender et al., 2021; Raji et al., 2021; Church & Kordoni, 2022; Hullman et al., 2022). Despite strong professional incentives encouraging this practice, it lacks scientific, engineering, and societal justification. Benchmarks poorly reflect model performance in real application contexts because of problems like data leakage, overfitting to benchmarks, and data heterogeneity (El-Mhamdi et al., 2021; 2023; Hanneke & Kpotufe, 2022; Balloccu et al., 2024; Xu et al., 2024; Zhang et al., 2024). In short, the measurement method lacks *external validity*. Yet the practice persists

[5] On similar lottery or path dependence effects, see Liebowitz & Margolis (1995); Peacock (2009); Dehghani et al. (2021); Fishman & Hancox-Li (2022); Rossbach (2023); Bauer & Gill (2024); Hooker (2024).

due to reputational and financial rewards, demonstrating misalignment between incentivized goals and their actual merits.

The dynamics of goal lotteries are also visible in the story of the multi-decade neglect of deep learning architectures. In this case, a research agenda was sidelined for reasons that eventually proved to be misguided from an engineering, scientific, or societal perspective (e.g., due to "gatekeeping" effects against less popular research agendas; see Siler et al., 2015). Meanwhile, the AI industry went all in on the expert systems "bubble" (Haigh, 2024). *Reductions in diversity within* contemporary AI research can be a sign that similar mistakes are at play (§2.6). Some recent initiatives to counter homogenization (Chollet et al., 2024) rely on operationalizing AGI through benchmarks.[6] In practice, they end up as yet another benchmark: incentivizing SOTA-chasing, supercharged by intense media and marketing attention (Jones, 2025). For this reason, we remain somewhat skeptical of whether approaches like ARC (Chollet, 2019) outweigh the negative consequences of news-cycle-accelerated SOTA-chasing.

2.5. Generality Debt

Relying on the generality or flexibility of tools to postpone crucial engineering, scientific, or societal decisions

AGI definitions differ on how much "generality" is desirable (Blili-Hamelin et al., 2024). This indicates a lack of clarity and consensus about the goals of AI research, forming a trap that (a) encourages suboptimal science/engineering practices (related to points made in 2.2); (b) suppresses important social/ethical questions about which research directions are worth pursuing. We term this trap "Generality Debt" to parallel technical debt (Sculley et al., 2014): it delays the work that needs to be done as part of AI research which, if left undone, takes more work to address in the future.

This trap includes the appeal to many different notions of generality at play in machine learning: (1) variety of tasks (Hernández-Orallo & Seán Ó hÉigeartaigh, 2018); (2) capability to be trained for "any task" vs. ability to perform many predefined tasks (Hernández-Orallo & Seán Ó hÉigeartaigh, 2018); (3) whether the task or data distribution the model is being evaluated on is "seen" or "unseen" (i.e., available, or not, to the model during its training phase) (Altmeyer et al., 2024); (4) variety of data in model input/output, such as structured vs unstructured, modality, etc.; (5) whether the performance of the model reflects "performance considered 'surprising' to humans" (Altmeyer et al., 2024); (6) variety of goals; (7) ability to "accept a general language for the problem statement" (Newell & Ernst, 1965); and (8) having a "general" internal representation (Newell & Ernst, 1965; McCarthy & Hayes, 1981).

As Paolo et al. (2024) note, despite the multiple possible meanings of "generality", most papers do not define generality even if it is central to their argument. Without formal definition, assessing or improving generalization becomes challenging. Assuming that "generalization" is desirable while acknowledging its poor definition is misguided. We should first define specific, measurable properties before arguing that they are desirable.

Without proper definition, the value of generality remains unclear. Different types of generality support different future visions, raising unexplored questions about their relative importance. Further, vague definitions of generality lead to bad science and engineering (§2.2). For example, Altmeyer et al. (2024) note how the pursuit of generality has led to vague task specifications. In parallel, Gebru & Torres (2024) argue that some conceptions of AGI contravene good engineering practices: it is hard to test the functionality of systems under "standard operating conditions" if the system is advertised as a "universal algorithm for learning and acting in any environment."

Aiming to achieve certain forms of generality could also mean making a tradeoff with ecological validity, as argued by Saxon et al. (2024). They argue that, in practice, "holistic" benchmarks tend to be a collection of disparate specific benchmark tasks, meaning that they have task-level construct validity. However, these tasks do not always match with *user-relevant capabilities*. Methodological challenges to achieving such capabilities may or may not be overcome, in time, given innovative solutions, but the pursuit of AGI assumes such challenges are surmountable. Methodological issues often inspire novel solutions, but we cannot assume a solution will be found, nor can those pursuing AGI. Further, strategies for pursuing AGI may introduce or reveal new challenges to achieving user-relevant capabilities.

Finally, the vagueness around "generality" is also an ethical concern, as it sidesteps normative questions about *which types of generality merit pursuit* and obscures implicit decisions about how to prioritize different research directions.

2.6. Normalized Exclusion

Excluding communities and experts from shaping goals

The negative consequences of exclusion in AI have been extensively discussed, both in terms of how it affects product quality and model performance (e.g., Obermeyer et al., 2019), and also how it harms people (e.g., Buolamwini & Gebru, 2018; Shelby et al., 2023; Whitney & Norman, 2024). We argue that AGI discourse

[6] Chollet proposes that "We will have AGI when creating [benchmarks 'that are easy for humans, yet impossible for AI'] becomes outright impossible" (2024b).

aggravates problems of exclusion.

Problem 1: Excluding communities. Many communities are left out of meaningful participation in shaping the goals of AI research (Delgado et al., 2023). Excluding communities causes serious harm, especially to minoritized communities (Pierre et al., 2021); it also undermines the utility of end products, reduces model performance, oversimplifies technical challenges (Kierans et al., 2025), and can impede innovation (e.g., Burt, 2004). For instance, the infamous case of facial recognition engines—such as those used by Google, Apple, and Meta—mistaking Black people for gorillas (BBC News, 2015) is still occurring more than 8 years after the problem was first identified (Appelman, 2023; Grant & Hill, 2023), with downstream impacts on surveillance and law enforcement (Jones, 2020; Pour, 2023). Similarly, selective forms of inclusion in data annotation raise ethical and practical concerns about downstream effects (Wang et al., 2022; Bertelsen et al., 2024). Other examples of exclusion or inclusion of communities impacting performance are plentiful (see Buolamwini & Gebru, 2018; Young et al., 2019; Raji et al., 2020; Andrews et al., 2024; Bergman et al., 2024; Salavati et al., 2024; Weidinger et al., 2024). Excluding communities from meaningful feedback also undermines societal goals, such as fostering collective legitimacy through accountability to impacted communities (Schulz et al., 2002; Mikesell et al., 2013; Costanza-Chock, 2020; Birhane et al., 2022a; Young et al., 2024).

The recent prominence of AGI discourse intensifies the existing problem of community exclusion in AI research (Frank et al., 2017; Kelly, 2024). Many proponents of AGI envision a future where AI systems perform an extraordinary range of tasks for countless communities. However, research and design processes fall short of the inclusiveness demanded by this ambitious vision. For example, December 2024 reporting suggests that OpenAI and Microsoft "signed an agreement last year stating OpenAI has only achieved AGI when it develops AI systems that can generate at least $100 billion in profits" (OpenAI, 2025d; Zeff, 2025), a stark departure from OpenAI's public definition of AGI (OpenAI, 2018). As several authors argue, economic value is not the only type of desirable value (Agrawal et al., 2022; Dulka, 2022; Harrigian et al., 2023; Morris et al., 2024; Pierson et al., 2025). The economic definition helps guide the engineering decisions of OpenAI. However, it is questionable whether an emphasis on profits will lead to the most beneficial or useful end products or to meaningful consensus about goals, especially for minoritized groups.

Problem 2: Excluding disciplines. From application domains (e.g., medicine (Obermeyer et al., 2019), finance (Cao, 2022), cybersecurity (Salem et al., 2024), learning (Leong & Linzen, 2024)) to the practices involved in building AI—data annotation, qualitative and quantitative methods, domain expertise, computer science, and many more (Wang et al., 2022; Bertelsen et al., 2024; Widder, 2024)—AI research crosses disciplinary boundaries. The cross-disciplinary challenges of AI research mirror those of other disciplines (Stokols et al., 2003; Stirling, 2014; Amoo et al., 2020; Vestal & Mesmer-Magnus, 2020; The Royal Society, 2024).

One major challenge is disciplinary silos, where knowledge is inadequately shared across disciplines (Stokols et al., 2003; Stirling, 2014; Ballantyne, 2019; Amoo et al., 2020; DiPaolo, 2022; The Royal Society, 2024). For instance, lack of knowledge sharing could be partly responsible for low attention to the distinction between explanatory and exploratory research in ML, discussed in §2.2 (Herrmann et al., 2024).

Another challenge is epistemic hierarchies—where the expertise of some disciplines is explicitly or implicitly devalued (Knorr Cetina, 1999; 2007; Simonton, 2004; Fourcade et al., 2015; Graziul et al., 2023). This can manifest as expert groups being limited to narrow input rather than contributing to broader research design decisions (Bertelsen et al., 2024).

AI researchers' focus on applying their work to other domains creates another major challenge. Insufficient domain knowledge might affect the functionality of AI tools—whether they operate as advertised (Raji et al., 2022). For instance, AI tools are deployed to make predictions about future individual-level outcomes, from pre-trial risk prediction and predictive policing to automated employment decisions (Wang et al., 2024). Yet inadequate evidence of effectiveness often fails to prevent predictive tools from being built, marketed, and deployed (Doucette et al., 2021; Cameron, 2023; Connealy et al., 2024).

The problem of disciplinary silos becomes particularly acute in AGI-oriented research due to two factors: its claims to be creating cognates or replacements of human intelligence, and its claims to expertise in many disciplines. In the former case, claims are often made while ignoring debates in cognitive science and psychology about the nature of intelligence (Summerfield, 2023; Guest & Martin, 2024; Mitchell, 2024; Van Rooij et al., 2024). In the latter case, achieving AGI is often framed in terms of being able to "replace" domain experts in various domains—which are then often taken up uncritically by the media without input from domain experts themselves (e.g., Henshall, 2024).

Problem 3: Resource disparities. In recent years, we have witnessed an unprecedented growth in computational resources required for model training, with requirements doubling approximately every few months (Sevilla et al., 2022). This trend compounds existing resource disparities, as state-of-the-art AI research often relies on access to computational resources accessible to very few researchers (Yu et al., 2023; LaForge, 2024). The financial cost of these resources excludes a wide range of actors from contributing to AI research, as even top research universities have a fraction of the computational resources that many corporations use to advance AI research. Efforts are underway to address this resource disparity by supporting access to large-scale computational resources maintained by government entities (e.g., NAIRR in the United States). Yet, resource parity is an aspirational goal in response to widespread recognition that AI researchers in industry enjoy a *de facto* advantage in setting the goals of AI research due to their access to industrial scale computational resources. This structural advantage is reinforced by the use of pre-print archives to publicize AI research without peer review (Devlin et al., 2019; Rombach et al., 2022; Bubeck et al., 2023; OpenAI et al., 2024), a strategy which legitimizes this work as scientific in nature without applying traditional standards for scientific integrity (Tenopir et al., 2016; Lin et al., 2020; Soderberg et al., 2020; Rastogi et al., 2022; Kwon & Porter, 2025).

While resource disparities exist for all forms of AI research, they are particularly stark when AGI is taken as a north-star goal for the discipline, due to the orientation of current AGI efforts towards sheer computational scale and the concentration of such efforts in large tech companies.[7] Such concentration of power makes it even more important that those efforts include, rather than exclude, relevant communities and experts. That is, AGI discourse accelerates the existing trend in AI of discounting domain expertise and lived experiences in favor of models that are allegedly experts in everything.

3. Recommendations

We have argued that AGI discourse hinders setting well-motivated scientific and engineering goals in AI development, while being destructive to the development of AI that has social merit. We now provide three recommendations for avoiding these traps.

Recommendation 1: Goal Specificity. *The AI community must prioritize **highly specific** language when discussing the scientific, engineering, and societal goals of AI.*

More specific definitions of tangible scientific, engineering, and societal goals promote a shared understanding of these goals, and thus the capacity to evaluate whether these goals are well-motivated. Without such specificity, researchers, practitioners, and others can develop divergent understandings of a goal and how it should be achieved. This divergence enables conceptual arbitrage on the part of AI researchers and practitioners who seek to advance their own goals, since these actors ultimately determine the details of model development and implementation. People external to AI development may then be left assuming that a system achieves a specific goal when, in fact, it does not.

Specificity can maintain sufficient flexibility for exploratory research by engaging in best practices around developing a research question. Consider the research goal "How can a mixture of experts (MoE) strategy improve performance of Whisper large-v3 in challenging speech domains?" which could cover various domains of speech or MoE strategies. A reformulated, more specific goal would be: "How can a MoE strategy, where experts are a series of models fine-tuned on short (<2s), medium (2–20s), and long utterances (>20s), help improve performance of Whisper large-v3 in a speech domain dominated by short utterances but also containing relatively long utterances?" Without additional specification, answering the first question provides little guarantee that a solution would address the specific features of the second question, at least not without substantial effort to understand how such a general solution may be applied to this specific case/domain. This example is illustrative, though based on real features of a speech domain where accuracy is essential (i.e., police radio communications, see Srivastava et al. 2024; Venkit et al. 2024).

Goal Specificity addresses the Illusion of Consensus trap (§2.1), promoting conceptual clarity as an essential part of goal-setting. Clarity also helps to avoid the Goal Lottery trap (§2.4) by making goal selection explicit. It similarly addresses the Presuming Value-Neutrality Trap (§2.3) by explicitly surfacing values tied to specific goals, and directly reduces Generality Debt (§2.5) . Finally, goal specificity addresses the underspecification issues highlighted in the Supercharging Bad Science Trap (§2.2).

Recommendation 2: Pluralism of goals and approaches. *Rather than a single general north-star goal (or small set of goals), the AI community should articulate **many** worthwhile scientific, engineering, and societal goals—and many possible paths to fulfilling them.*

Reaching meaningful scientific and societal consensus on the goals of a field as broad-ranging as AI is challenging. When consensus may not be viable or desirable, we recommend pluralism: allowing multiple viable conceptions of the goals of AI research. Pluralism is healthy in a society composed of individuals and institutions with divergent values. By default, the research community should be pluralistic about goals and paths to achieving them, aiming for heterogeneity instead of homogeneity (Sorensen et al.,

[7] Large-scale efforts also have detrimental impacts on climate, reinforcing resource disparities (Bucknall & Dori-Hacohen, 2022; Kaack et al., 2022; Luccioni et al., 2024).

2024a;b).[8]

Researchers who study the dynamics of knowledge production and problem-solving in groups have found pluralism to be beneficial (Hong & Page, 2004; Muldoon, 2013), including unique benefits ascribable to egalitarian group dynamics (Xu et al., 2022).

Pluralism in AI research can take several forms, each with distinct implications for how we approach complex problems. For example:

Methodological pluralism implies that different ways of approaching a problem help achieve better solutions, often an effective strategy for complex problem-solving in general (Midgley, 2000; Veit, 2020; Zhu, 2022) .

Value pluralism, as applied to alignment research, implies a direct connection between technical advancement and accommodation of plural values (Sorensen et al., 2024a;b).

Algorithmic pluralism addresses the “patterned inequality” associated with algorithmic monoculture and implies a “plurality of paths to different outcomes” must be supported to avoid reproducing existing social inequalities (e.g., embedded in data) (Jain et al., 2024) .

Each form of pluralism translates into concrete research practices, such as choice of method(s), setting of pluralistic goals, and testing algorithms to ensure known harms (i.e., algorithmic discrimination) are addressed.

To name just one example, algorithmic decision-making in hiring processes is unlikely to benefit from AGI as much as from targeted solutions to that particular setting, which account for the domain-specific nature of most positions, the different ways hiring managers evaluate candidates, and existing evidence of algorithmic discrimination in hiring. In practice, pluralism can manifest in how resources are distributed among different research approaches. For example, rather than investing most computational resources in pursuit of AGI, they could be more evenly distributed among diverse goals and approaches within AI.

Pluralism addresses the Illusion of Consensus trap (§2.1) by acknowledging the lack of consensus, and using diversity of perspectives as a tool for scientific and social progress; the Goal Lottery trap (§2.4) by reducing the chances of arbitrarily or prematurely excluding some goals from consideration; and the Exclusion trap (§2.6) by encouraging a plurality of goals and approaches.

[8] We’re not rejecting consensus on unifying, general north-star goals as a matter of principle. In some circumstances, like coordinating collective action in response to the climate crisis, strong consensus may become crucial. But the research community should not begin from the assumption that such consensus is necessary, or that consensus is optimal from a scientific or societal perspective.

Recommendation 3: Greater Inclusion in Goal Setting. *Greater inclusion of communities and disciplines in shaping the goals of AI research is beneficial to innovation.*

Inclusion supports innovation (Burt, 2004; Hewlett et al., 2013; Zhang et al., 2021; Xu et al., 2022). Identifying worthwhile goals, related use cases, and potential unintended consequences depends on engaging diverse viewpoints. Within AI research, these viewpoints must include those of end users, experts from other fields, those affected by research outcomes, and data annotators. Excluding any of these groups impoverishes the potential of AI to achieve worthwhile goals since it would discount the perspectives that define these goals as worthwhile. Including them enriches AI research.

Cross-pollination of ideas between disciplines leads to more impactful research (Dori-Hacohen et al., 2021; Shi & Evans, 2023). Such impact requires that we abandon silos of (tacit) knowledge (e.g., epistemic cultures: Knorr Cetina, 1999; 2007) and prioritize epistemic hierarchies that value non-computational research (Simonton, 2004; Fourcade et al., 2015). While technical complexity can make participation by non-experts challenging (Pierre et al., 2021), working through these challenges can surface issues experts have not anticipated (Cooper et al., 2022) and enable practical scientific contributions by integrating the insights and experiences of non-experts, or experts in other fields, into system design decisions (Delgado et al., 2023; Salavati et al., 2024).

As a topic, AGI often involves imagining AI technologies that impact the lives of everyone. Exclusion thus causes socially significant disagreements regarding the goals, processes, and actors who shape AI research and deployment. These disagreements are often overlooked or ignored by those with the power to shape the field. Inclusion is necessary to ensure that decisions about technology are sufficiently justified to institutions, communities, and individuals (Anderson, 2006; Binns, 2018; Alexandrova & Fabian, 2022; Birhane et al., 2022a; Lazar, 2022; Ovadya, 2023).

This recommendation addresses the Normalized Exclusion trap (§2.6) by treating inclusion as essential to innovation. Moreover, it addresses the Illusion of Consensus and the Presuming Value-Neutrality traps (§2.1, §2.3) by acknowledging socially significant disagreements about the value-laden goals of AI research and development and using those disagreements productively.

4. Alternative Views

We have argued that AGI is a poor choice as a north-star to guide AI research. We conclude by championing our position against a strong alternative view: that the traps we have identified can be addressed through a modified pursuit of AGI. We argue that improved approaches to AGI would not go far enough.

4.1. Counterargument

"*AGI is a good north-star goal; to avoid the above traps, improved definitions and accounts of AGI are needed.*"

Thoughtful attempts to address shortcomings in accounts of AGI indeed exist (e.g., Adams et al., 2012; Chollet, 2019; Summerfield, 2023; Morris et al., 2024). If prior accounts of AGI are counterproductive or flawed, why not pursue new accounts that address those flaws?

As an example of an alternative view, Morris et al. (2024) arguably mitigate the Illusion of Consensus trap by disentangling the disagreements about goals, predictions, and risks that plague other accounts of AGI. Moreover, their proposal for a practical strategy analogous to Levels of Driving Automation standards (SAE International, 2021) could be viewed as mitigating the Presuming Value-Neutrality trap. Setting society-wide standards could be done in a way that explicitly centers *specific* risks that the standards address (Morris et al., 2024). In this way, the values being favored manifest in the risks that are centered by the standards. Such works showcase a potential response to traps. In arguing that AGI discourse aggravates multiple standing problems, we cannot rule out the possibility of efforts that mitigate these same problems while retaining AGI as a goal. Moreover, given that lack of agreement about how to define AGI is likely to persist, as Summerfield (2023), Morris et al. (2024), and many others believe, it would be especially implausible for us to presume to have a complete enough view of the landscape of possible conceptions of AGI to draw definitive conclusions.

4.2. Rebuttal

Why favor our position against this alternative?

Reason 1: Conflict with our recommendations. Although the definition of AGI is highly contested, a frequent motivation for embracing AGI as a north-star goal is the desire for a single, large-scale, unifying vision for the field (Summerfield, 2023). This is somewhat in tension with our recommendation of goal pluralism, which we argue is valuable for AI research. "AGI" also brings with it a notion of "general" that can be discordant with our recommendation of specificity.[9] As such, there are reasons to be wary of AGI-focused alternatives.

[9]Note that these are not *all things considered* (or definitive) reasons to abandon the alternative view. Rather, readers who are convinced by our arguments in favor of specificity and pluralism have, to some extent, reasons to be wary of the project of improving AGI. These are so-called *pro-tanto* (i.e., "to that extent") reasons, which can be overridden by other considerations (Alvarez, 2023).

Reason 2: Distinguishing hype from reality. Another reason to favor our position is the AI research community's responsibility to help distinguish hype from reality. We believe that our community must provide trusted, evidence-based answers to increasingly complex questions about AI technologies, their goals, and their impacts. In the current moment, the hyped terminology of AGI undermines this responsibility.

No matter how well or poorly defined, AGI has acquired a cultural significance that exacerbates the challenge of distinguishing hype from reality. "Intelligence" and "generality" hold the promise of being beneficial for countless needs and contexts (§2.3, §2.5). No matter how cautious the research community attempts to be, the cultural associations of these terms risk stoking the flames of unscientific thinking about AI. This enables various parties to loosely project utopian or dystopian characteristics onto AGI in ways that support their calls for more power and resources.

Reason 3: Benefiting humans as the goal of technology. If the AI community nevertheless wants an overarching goal to strive towards, the goal should be the support and benefit of human beings. The goals of technology are shaped by *people*. Evidence-based approaches to examining whether technology effectively meets the needs of people—be they "users", "consumers", "patients", "scholars", or a myriad of business and social monikers—are well-established. In a quest to achieve AGI, communities often lose sight of the needs of people as a goal, in favor of focusing on just the technology.

There is another, more ambitious reason to work towards consensus on supporting and benefiting human beings as a goal. We have noted the role of socially significant disagreements about the goals of technology in our third recommendation of inclusion. Processes ensuring that technology benefits humans have the potential to provide *collectively legitimate* responses to such disagreements. This could occur through processes that embrace democratic ideals: such as universal inclusion in interrogation, deliberation, and dissent about the "common good" and "public interest", while enacting strong accountability to participants as "rights-holders" (Anderson, 2006; Putnam, 2011; Binns, 2018; Gabriel, 2020; Birhane et al., 2022a; Lazar & Nelson, 2023; Ovadya, 2023; Blili-Hamelin et al., 2024). Aiming for collective legitimacy amounts to requiring politically and socially effective forms of consensus.

In sum, we urge communities to **stop treating "AGI" as the north-star goal of AI research**.

Acknowledgments

Borhane Blili-Hamelin was funded in part through a Magic Grant from the Brown Institute for Media Innovation.

Christopher Graziul was supported by the National Institute of Minority Health and Health Disparities of the National Institutes of Health under award number R01MD015064. This material is based upon work supported in part by the NSF Program on Fairness in AI in Collaboration with Amazon under Award IIS-2147305. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, National Institutes of Health, the Brown Institute for Media Innovation, or Amazon.

References

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D. G. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 252–260, Barcelona Spain, January 2020. ACM. ISBN 978-1-4503-6936-7. doi: 10.1145/3351095.3372871. URL https://dl.acm.org/doi/10.1145/3351095.3372871.

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J. S., Samsonovich, A., Scheutz, M., Schlesinger, M., et al. Mapping the landscape of human-level artificial general intelligence. *AI magazine*, 33(1): 25–42, 2012.

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. Large language models are few-shot clinical information extractors. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL https://aclanthology.org/2022.emnlp-main.130/.

Agre, P. E. Toward a critical technical practice: Lessons learned in trying to reform AI. In *Social science, technical systems, and cooperative work*, pp. 131–157. Psychology Press, 2014.

Agüera y Arcas, B. and Norvig, P. Artificial General Intelligence Is Already Here, October 2023. URL https://www.noemamag.com/artificial-general-intelligence-is-already-here.

Ahmed, S., Jaźwińska, K., Ahlawat, A., Winecoff, A., and Wang, M. Field-building and the epistemic culture of AI safety. *First Monday*, April 2024. ISSN 1396-0466. doi: 20240428092345000. URL https://firstmonday.org/ojs/index.php/fm/article/view/13626.

Alexandrova, A. and Fabian, M. Democratising Measurement: or Why Thick Concepts Call for Coproduction. *European Journal for Philosophy of Science*, 12 (1):7, January 2022. ISSN 1879-4920. doi: 10.1007/s13194-021-00437-7. URL https://doi.org/10.1007/s13194-021-00437-7.

Allen, L., O’Connell, A., and Kiermer, V. How can we ensure visibility and diversity in research contributions? how the contributor role taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1):71–74, 2019. doi: https://doi.org/10.1002/leap.1210. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/leap.1210.

Altmeyer, P., Demetriou, A. M., Bartlett, A., and Liem, C. C. S. Position: Stop Making Unscientific AGI Performance Claims. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 1222–1242. PMLR, July 2024. URL https://proceedings.mlr.press/v235/altmeyer24a.html. ISSN: 2640-3498.

Alvarez, M. Reasons for Action: Justification, Motivation, Explanation. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

Amodei, D. Machines of loving grace, 2024. URL https://www.darioamodei.com/essay/machines-of-loving-grace. [Online; accessed 19-May-2025].

Amoo, M. E., Bringardner, J., Chen, J.-Y., Coyle, E. J., Finnegan, J., Kim, C. J., Koman, P. D., Lagoudas, M. Z., Llewellyn, D. C., Logan, L., et al. Breaking down the silos: Innovations for multidisciplinary programs. In *2020 ASEE Virtual Annual Conference Content Access*, 2020.

Anderson, E. Situated Knowledge and the Interplay of Value Judgments and Evidence in Scientific Inquiry. In Gärdenfors, P., Woleński, J., and Kijania-Placek, K. (eds.), *In the Scope of Logic, Methodology and Philosophy of Science: Volume Two of the 11th International Congress of Logic, Methodology and Philosophy of Science, Cracow, August 1999*, Synthese Library, pp. 497–517. Springer Netherlands, Dordrecht, 2002. ISBN 978-94-017-0475-5. doi: 10.1007/978-94-017-0475-5_8. URL https://doi.org/10.1007/978-94-017-0475-5_8.

Anderson, E. The Epistemology of Democracy. *Episteme*, 3(1-2):8–22, June 2006. ISSN 1750-0117, 1742-3600. doi: 10.3366/epi.2006.3.1-2.8. URL https://www.cambridge.org/core/journals/episteme/article/

abs/epistemology-of-democracy/F86F1D124D2E081116611043BD54CBD9.

Andrews, M., Smart, A., and Birhane, A. The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, 0(0), August 2024. ISSN 2666-3899. doi: 10.1016/j.patter.2024.101027. URL https://www.cell.com/patterns/abstract/S2666-3899(24)00160-0.

Anthropic. Anthropic's recommendations to OSTP for the U.S. AI Action Plan, 2025. URL https://www.anthropic.com/news/anthropic-s-recommendations-ostp-u-s-ai-action-plan. [Online; accessed 19-May-2025].

Appelman, N. Racist Technology in Action: Image recognition is still not capable of differentiating gorillas from Black people — racismandtechnology.center. https://racismandtechnology.center/2023/06/09/racist-technology-in-action-image-recognition-is-still-not-capable-of-differentiating-gorillas-from-black-people/, 2023. [Accessed 24-01-2025].

Attard-Frost, B. Queering intelligence: A theory of intelligence as performance and a critique of individual and artificial intelligence. In *Queer Reflections on AI*. Routledge, 2023. ISBN 978-1-00-335795-7. URL https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003357957-3/queering-intelligence-blair-attard-frost.

Ballantyne, N. Epistemic Trespassing. *Mind*, 128(510): 367–395, April 2019. ISSN 0026-4423. doi: 10.1093/mind/fzx042. URL https://doi.org/10.1093/mind/fzx042.

Balloccu, S., Schmidtová, P., Lango, M., and Dusek, O. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Graham, Y. and Purver, M. (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 67–93, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.5/.

Bauer, K. and Gill, A. Mirror, Mirror on the Wall: Algorithmic Assessments, Transparency, and Self-Fulfilling Prophecies. *Information Systems Research*, 35(1):226–248, March 2024. ISSN 1047-7047. doi: 10.1287/isre.2023.1217. URL https://pubsonline.informs.org/doi/full/10.1287/isre.2023.1217.

BBC News. Google apologises for Photos app's racist blunder. https://www.bbc.com/news/technology-33347866, 2015. [Accessed 24-01-2025].

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. 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, pp. 610–623, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Bergman, S., Marchal, N., Mellor, J., Mohamed, S., Gabriel, I., and Isaac, W. STELA: a community-centred approach to norm elicitation for AI alignment. *Scientific Reports*, 14(1):6616, March 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-56648-4. URL https://www.nature.com/articles/s41598-024-56648-4. Publisher: Nature Publishing Group.

Bertelsen, P. S., Bossen, C., Knudsen, C., and Pedersen, A. M. Data work and practices in healthcare: A scoping review. *International Journal of Medical Informatics*, 184:105348, April 2024. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2024.105348. URL https://www.sciencedirect.com/science/article/pii/S138650562400011X.

Binns, R. Algorithmic Accountability and Public Reason. *Philosophy & Technology*, 31(4):543–556, December 2018. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-017-0263-5. URL http://link.springer.com/10.1007/s13347-017-0263-5.

Birhane, A. and Guest, O. Towards Decolonising Computational Sciences. *Kvinder, Køn & Forskning*, 2021. doi: 10.7146/kkf.v29i2.124899. URL https://pure.mpg.de/rest/items/item_3287104_1/component/file_3287105/content.

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., and Mohamed, S. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, pp. 1–8, New York, NY, USA, October 2022a. Association for Computing Machinery. ISBN 978-1-4503-9477-2. doi: 10.1145/3551624.3555290. URL https://dl.acm.org/doi/10.1145/3551624.3555290.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 173–184, Seoul Republic of Korea, June 2022b. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533083. URL https://dl.acm.org/doi/10.1145/3531146.3533083.

Blili-Hamelin, B. and Hancox-Li, L. Making Intelligence: Ethical Values in IQ and ML Benchmarks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 271–284, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3593996. URL https://dl.acm.org/doi/10.1145/3593013.3593996.

Blili-Hamelin, B., Hancox-Li, L., and Smart, A. Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7:141–155, October 2024. URL https://ojs.aaai.org/index.php/AIES/article/view/31625.

Blodgett, S. L., Barocas, S., Daumé Iii, H., and Wallach, H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://www.aclweb.org/anthology/2020.acl-main.485.

Bommasani, R. Evaluation for Change. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8227–8239, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.522. URL https://aclanthology.org/2023.findings-acl.522/.

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., and Liang, P. S. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, 35:3663–3678, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html.

Bommasani, R., Singer, S., Appel, R. E., Cen, S., Cooper, A. F., Cryst, E., Gailmard, L. A., Gonzalez, J. E., Ho, D. E., Klaus, I., Lee, M. M., Liang, P., Reuel, A., Song, D., Spence, D., Wan, A., Wang, A., Zhang, D., Zittrain, J., Tour Chayes, J., Cuéllar, M.-F., and Fei-Fei, L. DRAFT REPORT of the Joint California Policy Working Group on AI Frontier Models. Technical report, Joint California Policy Working Group on AI Frontier Models, March 2025. URL https://www.cafrontieraigov.org/wp-content/uploads/2025/03/Draft_Report_of_the_Joint_California_Policy_Working_Group_on_AI_Frontier_Models.pdf.

Bostrom, N. *Superintelligence: Paths, dangers, strategies.* Oxford University Press, New York, NY, US, 2014. ISBN 978-0-19-967811-2.

Bouthillier, X., Laurent, C., and Vincent, P. Unreproducible research is reproducible. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 725–734. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/bouthillier19a.html.

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., and Chuan, C.-H. Artificial Intelligence and Journalism. *Journalism & Mass Communication Quarterly*, 96(3):673–695, September 2019. ISSN 1077-6990, 2161-430X. doi: 10.1177/1077699019859901. URL http://journals.sagepub.com/doi/10.1177/1077699019859901.

Browne, R. AI that can match humans at any task will be here in five to 10 years, Google DeepMind CEO says. https://www.cnbc.com/2025/03/17/human-level-ai-will-be-here-in-5-to-10-years-deepmind-ceo-says.html, 2025.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with GPT-4, 2023. URL https://arxiv.org/abs/2303.12712.

Bucknall, B. S. and Dori-Hacohen, S. Current and Near-Term AI as a Potential Existential Risk Factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 119–129, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9247-1. doi: 10.1145/3514094.3534146. URL https://dl.acm.org/doi/10.1145/3514094.3534146.

Buolamwini, J. and Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st*

Conference on Fairness, Accountability and Transparency, pp. 77–91. PMLR, January 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Burt, R. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–399, September 2004. ISSN 0002-9602. doi: 10.1086/421787. URL https://www.journals.uchicago.edu/doi/full/10.1086/421787. Publisher: The University of Chicago Press.

Calin-Jageman, R. J. and Cumming, G. The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, 73(sup1): 271–280, 2019.

Cameron, D. US Justice Department Urged to Investigate Gunshot Detector Purchases. *Wired*, September 2023. ISSN 1059-1028. URL https://www.wired.com/story/shotspotter-doj-letter-epic/.

Cao, L. Ai in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys*, 55(3), February 2022. ISSN 0360-0300. doi: 10.1145/3502289. URL https://doi.org/10.1145/3502289.

Cave, S. The Problem with Intelligence: Its Value-Laden History and the Future of AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’20, pp. 29–35, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375813. URL https://dl.acm.org/doi/10.1145/3375627.3375813.

Chalmers, D. J. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10):9 – 10, 2010.

Chollet, F. On the Measure of Intelligence, November 2019. URL http://arxiv.org/abs/1911.01547.

Chollet, F. OpenAI o3 Breakthrough High Score on ARC-AGI-Pub, December 2024a. URL https://arcprize.org/blog/oai-o3-pub-breakthrough.

Chollet, F. ”So, is this AGI?...”, December 2024b. URL https://x.com/fchollet/status/1870170778458828851.

Chollet, F., Knoop, M., Kamradt, G., and Landers, B. ARC Prize 2024: Technical Report. Technical report, ARC-AGI, December 2024.

Church, K. W. and Kordoni, V. Emerging Trends: SOTA-Chasing. *Natural Language Engineering*, 28(2):249–269, March 2022. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324922000043. URL https://www.cambridge.org/core/product/identifier/S1351324922000043/type/journal_article.

Connealy, N. T., Piza, E. L., Arietti, R. A., Mohler, G. O., and Carter, J. G. Staggered deployment of gunshot detection technology in Chicago, IL: a matched quasi-experiment of gun violence outcomes. *Journal of Experimental Criminology*, March 2024. ISSN 1572-8315. doi: 10.1007/s11292-024-09617-w. URL https://doi.org/10.1007/s11292-024-09617-w.

Cooper, N., Horne, T., Hayes, G. R., Heldreth, C., Lahav, M., Holbrook, J., and Wilcox, L. A Systematic Review and Thematic Analysis of Community-Collaborative Approaches to Computing Research. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, pp. 1–18, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3517716. URL https://dl.acm.org/doi/10.1145/3491102.3517716.

Costanza-Chock, S. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press, 2020. ISBN 978-0-262-04345-8. URL https://library.oapen.org/handle/20.500.12657/43542.

D’Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL http://jmlr.org/papers/v23/20-1335.html.

DeepMind, G. About, 2025. URL https://deepmind.google/about/. [Online; accessed 19-May-2025].

Dehghani, M., Tay, Y., Gritsenko, A. A., Zhao, Z., Houlsby, N., Diaz, F., Metzler, D., and Vinyals, O. The Benchmark Lottery, July 2021. URL http://arxiv.org/abs/2107.07002.

Delgado, F., Yang, S., Madaio, M., and Yang, Q. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–23, Boston MA USA, October 2023.

ACM. ISBN 9798400703812. doi: 10.1145/3617694.3623261. URL https://dl.acm.org/doi/10.1145/3617694.3623261.

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M. K. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets, July 2020. URL http://arxiv.org/abs/2007.07399. arXiv:2007.07399 [cs].

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

DiPaolo, J. What's wrong with epistemic trespassing? *Philosophical Studies*, 179(1):223–243, January 2022. ISSN 1573-0883. doi: 10.1007/s11098-021-01657-6. URL https://doi.org/10.1007/s11098-021-01657-6.

Dori-Hacohen, S., Montenegro, R. E., Murai, F., Hale, S. A., Sung, K., Blain, M., and Edwards-Johnson, J. Fairness via AI: Bias Reduction in Medical Information. In *The 4th FAccTRec Workshop on Responsible Recommendation at RecSys*, 2021.

Dotan, R. and Milli, S. Value-laden disciplinary shifts in machine learning | Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020. doi: 10.1145/3351095.3373157. URL https://dl.acm.org/doi/abs/10.1145/3351095.3373157.

Doucette, M. L., Green, C., Necci Dineen, J., Shapiro, D., and Raissian, K. M. Impact of ShotSpotter Technology on Firearm Homicides and Arrests Among Large Metropolitan Counties: a Longitudinal Analysis, 1999–2016. *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, 98(5):609–621, October 2021. ISSN 1099-3460. doi: 10.1007/s11524-021-00515-4. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8566613/.

Dreyfus, H. and Dreyfus, S. E. *Mind Over Machine*. Simon and Schuster, 1986. ISBN 978-0-7432-0551-1.

Dulka, A. The Use of Artificial Intelligence in International Human Rights Law. *Stanford Technology Law Review*, 26(2):316–366, 2022. URL https://heinonline.org/HOL/P?h=hein.journals/stantlr26&i=316.

El-Mhamdi, E. M., Farhadkhani, S., Guerraoui, R., Guirguis, A., Hoang, L.-N., and Rouault, S. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in neural information processing systems*, 34: 25044–25057, 2021.

El-Mhamdi, E.-M., Farhadkhani, S., Guerraoui, R., Gupta, N., Hoang, L.-N., Pinot, R., Rouault, S., and Stephan, J. On the Impossible Safety of Large AI Models, May 2023. URL http://arxiv.org/abs/2209.15259. arXiv:2209.15259 [cs].

Fast Company. Wozniak: Could a computer make a cup of coffee?, 2010. URL https://www.youtube.com/watch?v=MowergwQR5Y. [Online; accessed 17-January-2023].

Fei, N., Lu, Z., Gao, Y., Yang, G., Huo, Y., Wen, J., Lu, H., Song, R., Gao, X., Xiang, T., Sun, H., and Wen, J.-R. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13 (1):3094, June 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-30761-2. URL https://www.nature.com/articles/s41467-022-30761-2. Publisher: Nature Publishing Group.

Fishman, N. and Hancox-Li, L. Should attention be all we need? The epistemic and ethical implications of unification in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1516–1527, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533206. URL https://dl.acm.org/doi/10.1145/3531146.3533206.

Fourcade, M., Ollion, E., and Algan, Y. The Superiority of Economists. *Journal of Economic Perspectives*, 29(1):89–114, February 2015. ISSN 0895-3309. doi: 10.1257/jep.29.1.89. URL https://www.aeaweb.org/articles?id=10.1257/jep.29.1.89.

Francesca Rossi, Christian Bessiere, Joydeep Biswas, Rodney Brooks, Vincent Conitzer, Thomas G. Dietterich, Virginia Dignum, Oren Etzioni, Kenneth D. Forbus, Eugene Freuder, Yolanda Gil, Holger Hoos, Eric Horvitz, Subbarao Kambhampati, Henry Kautz, Jihie Kim, Hiroaki Kitano, Alan Mackworth, Karen Myers, Luc De Raedt, Stuart Russell, Bart Selman, Peter Stone, Millind Tambe, and Michael Wooldridge. AAAI 2025 presidential panel on the future of AI research. Technical report, Association for the Advancement of Artificial Intelligence, March 2025. URL https://aaai.org/wp-content/uploads/2025/03/AAAI-2025-PresPanel-Report-Digital-3.7.25.pdf.

Frank, M., Roehrig, P., and Pring, B. *What to do when machines do everything: How to get ahead in a world of AI, algorithms, bots, and big data*. John Wiley & Sons, 2017.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, April 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3433949. URL https://dl.acm.org/doi/10.1145/3433949.

Gabriel, I. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-020-09539-2. URL https://link.springer.com/10.1007/s11023-020-09539-2.

Gebru, T. and Torres, E. P. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, April 2024. ISSN 1396-0466. doi: 20240428092319000. URL https://firstmonday.org/ojs/index.php/fm/article/view/13636.

Goertzel, B. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014. URL https://sciendo.com/abstract/journals/jagi/5/1/article-p1.xml.

Goertzel, B., Iklé, M., and Wigmore, J. The architecture of human-like general intelligence. In *Theoretical foundations of artificial general intelligence*, pp. 123–144. Springer, 2012.

Gopnik, A. AIs Versus Four-Year-Olds. In Brockman, J. (ed.), *Possible minds: twenty-five ways of looking at AI*. Penguin Press, New York, 2019. ISBN 978-0-525-55799-9 978-0-525-55801-9.

Gould, S. J. *The mismeasure of man*. Norton, New York, 1st ed edition, 1981. ISBN 978-0-393-01489-1.

Grant, N. and Hill, K. Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's. (Published 2023) — nytimes.com. https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html, 2023. [Accessed 24-01-2025].

Graziul, C., Belikov, A., Chattopadyay, I., Chen, Z., Fang, H., Girdhar, A., Jia, X., Krafft, P. M., Kleiman-Weiner, M., Lewis, C., Liang, C., Muchovej, J., Vientós, A., Young, M., and Evans, J. Does big data serve policy? Not without context. An experiment with in silico social science. *Computational and Mathematical Organization Theory*, 29(1):188–219, March 2023. ISSN 1572-9346. doi: 10.1007/s10588-022-09362-3. URL https://doi.org/10.1007/s10588-022-09362-3.

Green, B. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing*, 2(3):249–265, September 2021. ISSN 2688-5255. doi: 10.23919/JSC.2021.0029. URL https://ieeexplore.ieee.org/abstract/document/9684742.

Grossman, G. AGI is coming faster than we think: We must get ready now. *VentureBeat*, 2023. URL https://venturebeat.com/ai/agi-is-coming-faster-than-we-think-we-must-get-ready-now/. Accessed: Jan 17, 2025.

Gruetzemacher, R. and Whittlestone, J. The transformative potential of artificial intelligence. *Futures*, 135:102884, January 2022. ISSN 00163287. doi: 10.1016/j.futures.2021.102884. URL https://linkinghub.elsevier.com/retrieve/pii/S0016328721001932.

Gubrud, M. A. Nanotechnology and International Security. In *Fifth Foresight Conference on Molecular Nanotechnology*, volume 1, 1997. URL https://web.archive.org/web/20110529215447/http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/.

Guest, O. and Martin, A. E. A Metatheory of Classical and Modern Connectionism, October 2024. URL https://osf.io/eaf2z.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Haigh, T. How the AI boom went bust. *Commun. ACM*, 67(2):22–26, January 2024. ISSN 0001-0782. doi: 10.1145/3634901. URL https://doi.org/10.1145/3634901.

Hanneke, S. and Kpotufe, S. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

Hao, K. The democracy summit 2023, 2023. URL https://www.youtube.com/live/0fkGiZ0WqRc?si=NZ9hdvOQLcHNyC4Q&t=28498. [Panel video online; accessed 17-January-2025].

Harrigian, K., Zirikly, A., Chee, B., Ahmad, A., Links, A., Saha, S., Beach, M. C., and Dredze, M. Characterization of Stigmatizing Language in Medical Records. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 312–329, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.28. URL https://aclanthology.org/2023.acl-short.28/.

Henshall, W. When Might AI Outsmart Us? It Depends Who You Ask. *Time*, January 2024. URL https://time.com/6556168/when-ai-outsmart-humans/.

Hernández-Orallo, J. and Seán Ó hÉigeartaigh, S. Paradigms of artificial general intelligence and their associated risks. *Centre for the Study of Existential Risk, University of Cambridge, UK*, 2018.

Hernández-Orallo, J., Dowe, D. L., and Hernández-Lloreda, M. Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research*, 27:50–74, 2014. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2013.06.001. URL https://www.sciencedirect.com/science/article/pii/S1389041713000338.

Hernández-Orallo, J., Loe, B. S., Cheke, L., Martínez-Plumed, F., and Ó hÉigeartaigh, S. General intelligence disentangled via a generality metric for natural and artificial intelligence. *Scientific Reports*, 11(1):22822, November 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-01997-7. URL https://www.nature.com/articles/s41598-021-01997-7.

Herrmann, M., Lange, F. J. D., Eggensperger, K., Casalicchio, G., Wever, M., Feurer, M., Rügamer, D., Hüllermeier, E., Boulesteix, A.-L., and Bischl, B. Position: Why We Must Rethink Empirical Research in Machine Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 18228–18247. PMLR, July 2024. URL https://proceedings.mlr.press/v235/herrmann24b.html. ISSN: 2640-3498.

Hewlett, S. A., Marshall, M., and Sherbin, L. How Diversity Can Drive Innovation. *Harvard Business Review*, 91 (12), December 2013. ISSN 0017-8012.

Hicks, M. T., Humphries, J., and Slater, J. ChatGPT is bullshit. *Ethics and Information Technology*, 26(2): 38, June 2024. ISSN 1572-8439. doi: 10.1007/s10676-024-09775-5. URL https://doi.org/10.1007/s10676-024-09775-5.

Holland, S. Trump to announce private sector AI infrastructure investment, CBS reports. *Reuters*, January 2025. URL https://www.reuters.com/technology/artificial-intelligence/trump-announce-private-sector-ai-infrastructure-investment-cbs-reports-2025-01-21/.

Hong, L. and Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, November 2004. doi: 10.1073/pnas.0403723101. URL https://www.pnas.org/doi/full/10.1073/pnas.0403723101.

Hooker, S. The hardware lottery. *Commun. ACM*, 64 (12):58–65, November 2021. ISSN 0001-0782. doi: 10.1145/3467017. URL https://doi.org/10.1145/3467017.

Hooker, S. On the diminishing returns to scaling. [Online - Accessed 2024-01-12], Nov 2024. URL https://drive.google.com/file/d/1yeW429nx_FXaK_RgqDv89wH4Gh5flIRG/view.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., and Narayanan, A. 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, pp. 335–348, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534196. URL https://doi.org/10.1145/3514094.3534196.

Hutchinson, B., Rostamzadeh, N., Greer, C., Heller, K., and Prabhakaran, V. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1859–1876, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533233. URL https://doi.org/10.1145/3531146.3533233.

IBM. Getting ready for artificial general intelligence with examples, 2023. URL https://www.ibm.com/think/topics/artificial-general-intelligence-examples. Accessed: Jan 17, 2025.

Jacobs, A. Z. and Wallach, H. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445901. URL https://dl.acm.org/doi/10.1145/3442188.3445901.

Jain, S., Suriyakumar, V., Creel, K., and Wilson, A. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 197–206, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658899. URL https://dl.acm.org/doi/10.1145/3630106.3658899.

Jones, C. Law enforcement use of facial recognition: bias, disparate impacts on people of color, and the need for federal legislation. *NCJL & Tech.*, 22:777, 2020.

Jones, N. How should we test AI for human-level intelligence? OpenAI's o3 electrifies quest. *Nature*, 637 (8047):774–775, January 2025. ISSN 1476-4687. doi: 10.1038/d41586-025-00110-6. URL https://www.nature.com/articles/d41586-025-00110-6.

Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., and Rolnick, D. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change*, 12(6):518–527, 2022.

Kelly, S. M. Elon Musk says AI will take your job, and 'no one is going to need to work'. *CNN*, May 2024. URL https://www.cnn.com/2024/05/23/tech/elon-musk-ai-your-job/index.html. Accessed: Janury 19, 2025.

Kerner, S. M. Elon Musk reveals xAI efforts, predicts full AGI by 2029, 2023. URL https://venturebeat.com/ai/elon-musk-reveals-xai-efforts-predicts-full-agi-by-2029/. [Online; accessed 19-May-2025].

Kierans, A., Ghosh, A., Hazan, H., and Dori-Hacohen, S. Quantifying misalignment between agents: Towards a sociotechnical understanding of alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, March 2025. URL https://arxiv.org/abs/2406.04231.

Klein, E. Opinion | The Government Knows A.G.I. Is Coming. *The New York Times*, March 2025. ISSN 0362-4331. URL https://www.nytimes.com/2025/03/04/opinion/ezra-klein-podcast-ben-buchanan.html.

Kleinberg, J. and Raghavan, M. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2018340118.

Knorr Cetina, K. *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press, May 1999. ISBN 978-0-674-03968-1.

Knorr Cetina, K. Culture in global knowledge societies: knowledge cultures and epistemic cultures. *Interdisciplinary Science Reviews*, 32(4): 361–375, December 2007. ISSN 0308-0188. doi: 10.1179/030801807X163571. URL https://journals.sagepub.com/doi/abs/10.1179/030801807X163571.

Kwon, S. and Porter, A. L. Use of exclusive data for corporate research on machine learning and artificial intelligence: Implications for innovation and competition policy. *Technology in Society*, 81:102820, June 2025. ISSN 0160-791X. doi: 10.1016/j.techsoc.2025.102820. URL https://www.sciencedirect.com/science/article/pii/S0160791X25000107.

LaForge, G. The Dangers of Imposing Global North Approaches to AI Governance on the Global South | TechPolicy.Press, September 2024. URL https://techpolicy.press/the-dangers-of-imposing-global-north-approaches-to-ai-governance-on-the-global-south/.

Lazar, S. Power and AI: Nature and Justification. In Bullock, J., Chen, Y.-C., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M., and Zhang, B. (eds.), *The Oxford Handbook of AI Governance*. Oxford University Press, May 2022. ISBN 978-0-19-757932-9. doi: 10.1093/oxfordhb/9780197579329.013.12. URL https://oxfordhandbooks.com/view/10.1093/oxfordhb/9780197579329.001.0001/oxfordhb-9780197579329-e-12.

Lazar, S. and Nelson, A. AI safety on whose terms? *Science*, 381(6654):138–138, July 2023. doi: 10.1126/science.adi8982. URL https://www.science.org/doi/10.1126/science.adi8982.

Legg, S. and Hutter, M. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, 17 (4):391–444, December 2007. ISSN 1572-8641. doi: 10.1007/s11023-007-9079-x. URL https://doi.org/10.1007/s11023-007-9079-x.

Leong, C. S.-Y. and Linzen, T. Testing learning hypotheses using neural networks by manipulating learning data, July 2024. URL http://arxiv.org/abs/2407.04593. arXiv:2407.04593 [cs].

Liao, T., Taori, R., Raji, D., and Schmidt, L. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/757b505cfd34c64c85ca5b5690ee5293-Paper-round2.pdf.

Liebowitz, S. J. and Margolis, S. E. Path Dependence, Lock-in, and History. *Journal of Law, Economics, & Organization*, 11(1):205–226, 1995. ISSN 87566222, 14657341. URL http://www.jstor.org/stable/765077.

Lin, J., Yu, Y., Zhou, Y., Zhou, Z., and Shi, X. How many preprints have actually been printed and why: a case study of computer science preprints on arXiv. *Scientometrics*, 124(1):555–574, July 2020. ISSN 1588-2861. doi: 10.1007/s11192-020-03430-8. URL https://doi.org/10.1007/s11192-020-03430-8.

Luccioni, S., Gamazaychikov, B., Hooker, S., Pierrard, R., Strubell, E., Jernite, Y., and Wu, C.-J. Light bulbs have energy ratings—so why can't AI chatbots? *Nature*, 632 (8026):736–738, 2024.

Marcus, G. Dear Elon Musk, here are five things you might want to consider about AGI, 2022. URL https://garymarcus.substack.com/p/dear-elon-musk-here-are-five-things. [Online; accessed 24-January-2024].

Mathur, V., Lustig, C., and Kaziunas, E. Disordering Datasets: Sociotechnical Misalignments in AI-Mediated Behavioral Health. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–33, November 2022. ISSN 2573-0142. doi: 10.1145/3555141. URL https://dl.acm.org/doi/10.1145/3555141.

Maymin, P. Artificial general intelligence (AGI) is one prompt away, 2023. URL https://www.forbes.com/sites/philipmaymin/2023/10/13/artificial-general-intelligence-agi-is-one-prompt-away/. [Online; accessed 19-May-2025].

McCarthy, J. and Hayes, P. J. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pp. 431–450. Elsevier, 1981.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 1955. URL http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

Midgley, G. Methodological Pluralism. In Minati, G., Giuliani, A., and Bich, L. (eds.), *Systemic Intervention: Philosophy, Methodology, and Practice*, Contemporary Systems Thinking, pp. 171–216. Springer US, Boston, MA, 2000. ISBN 978-1-4615-4201-8. doi: 10.1007/978-1-4615-4201-8_9. URL https://doi.org/10.1007/978-1-4615-4201-8_9.

Mikesell, L., Bromley, E., and Khodyakov, D. Ethical Community-Engaged Research: A Literature Review. *American Journal of Public Health*, 103 (12):e7–e14, December 2013. ISSN 0090-0036. doi: 10.2105/AJPH.2013.301605. URL https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2013.301605.

Mitchell, M. Debates on the nature of artificial general intelligence. *Science*, 383(6689):eado7069, March 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ado7069. URL https://www.science.org/doi/10.1126/science.ado7069.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. Position: levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 36308–36321, Vienna, Austria, July 2024. JMLR.org.

Mueller, M. The myth of AGI. *Internet Governance Project*, 2024. URL https://www.internetgovernance.org/wp-content/uploads/MythofAGI.pdf.

Muldoon, R. Diversity and the Division of Cognitive Labor. *Philosophy Compass*, 8(2):117–125, 2013. ISSN 1747-9991. doi: 10.1111/phc3.12000. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12000.

Mulligan, D. K., Koopman, C., and Doty, N. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and*

Engineering Sciences, 374(2083):20160118, December 2016. doi: 10.1098/rsta.2016.0118. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2016.0118. Publisher: Royal Society.

Narayanan, A. and Kapoor, S. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can’t, and How to Tell the Difference*. Princeton University Press, September 2024. ISBN 978-0-691-24964-3. doi: 10.1515/9780691249643. URL https://www.degruyter.com/document/doi/10.1515/9780691249643/html.

Newell, A. and Ernst, G. The search for generality. In *Proc. IFIP Congress*, volume 65, pp. 17–24, 1965.

Nilsson, N. J. Human-Level Artificial Intelligence? Be Serious! *AI Magazine*, 26(4):68–68, December 2005. ISSN 2371-9621. doi: 10.1609/aimag.v26i4.1850. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1850.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366 (6464):447–453, October 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/full/10.1126/science.aax2342. Publisher: American Association for the Advancement of Science.

OpenAI. OpenAI Charter. Technical report, OpenAI, April 2018. URL https://openai.com/charter.

OpenAI. About, 2025a. URL https://openai.com/about/. [Online; accessed 19-May-2025].

OpenAI. Planning for AGI and beyond, 2025b. URL https://openai.com/index/planning-for-agi-and-beyond/. [Online; accessed 19-May-2025].

OpenAI. Security on the path to AGI, 2025c. URL https://openai.com/index/security-on-the-path-to-agi/. [Online; accessed 19-May-2025].

OpenAI. Our structure, 2025d. URL https://openai.com/our-structure/. [Online; accessed 17-January-2025].

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ovadya, A. Reimagining Democracy for AI. *Journal of Democracy*, 34(4):162–170, 2023. ISSN 1086-3214. doi: 10.1353/jod.2023.a907697. URL https://muse.jhu.edu/pub/1/article/907697.

Paolo, G., Gonzalez-Billandon, J., and Kégl, B. Position: A call for embodied AI. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39493–39508. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/paolo24a.html.

Peacock, M. S. Path Dependence in the Production of Scientific Knowledge. *Social Epistemology*, 23(2): 105–124, April 2009. ISSN 0269-1728, 1464-5297. doi: 10.1080/02691720902962813. URL http://www.tandfonline.com/doi/abs/10.1080/02691720902962813.

Perrigo, B. Meta’s AI chief Yann LeCun on AGI, open-source, and AI risk, 2024. URL https://time.com/6694432/yann-lecun-meta-ai-interview/. [Online; accessed 19-May-2025].

Pierre, J., Crooks, R., Currie, M., Paris, B., and Pasquetto, I. Getting Ourselves Together: Data-centered participatory design research & epistemic burden. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, pp. 1–11, New York, NY, USA, May 2021. Association for Computing Machinery. ISBN 978-1-4503-8096-6. doi: 10.1145/3411764.3445103. URL https://dl.acm.org/doi/10.1145/3411764.3445103.

Pierson, E., Shanmugam, D., Movva, R., Kleinberg, J., Agrawal, M., Dredze, M., Ferryman, K., Gichoya, J. W., Jurafsky, D., Koh, P. W., Levy, K., Mullainathan, S., Obermeyer, Z., Suresh, H., and Vafa, K. Using Large Language Models to Promote Health Equity. *NEJM AI*, 2(2):AIp2400889, January 2025. doi: 10.1056/AIp2400889. URL https://ai.nejm.org/doi/full/10.1056/AIp2400889. Publisher: Massachusetts Medical Society.

Pour, S. Police use of facial recognition technology and racial bias–an assessment of criticisms of its current use. *American Journal of Artificial Intelligence*, 7(1):17–23, 2023.

Putnam, H. A Reconsideration of Deweyan Democracy (Reprint from 1989). In *The pragmatism reader: from Peirce through the present*. Princeton University Press, Princeton, NJ Oxford, 2011. ISBN 978-0-691-13705-6 978-0-691-13706-3.

Raji, D., Denton, E., Bender, E. M., Hanna, A., and Paullada, A. AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’20, pp. 145–151, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375820. URL https://doi.org/10.1145/3375627.3375820.

Raji, I. D., Kumar, I. E., Horowitz, A., and Selbst, A. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, pp. 959–972, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533158. URL https://dl.acm.org/doi/10.1145/3531146.3533158.

Rastogi, C., Stelmakh, I., Shen, X., Meila, M., Echenique, F., Chawla, S., and Shah, N. B. To ArXiv or not to ArXiv: A Study Quantifying Pros and Cons of Posting Preprints Online, June 2022. URL http://arxiv.org/abs/2203.17259. arXiv:2203.17259 [cs].

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Rossbach, N. Innocent until Predicted Guilty: How Premature Predictive Policing Can Lead to a Self-Fulfilling Prophecy of Juvenile Delinquency Note. *Florida Law Review*, 75(1):167–194, 2023. URL https://heinonline.org/HOL/P?h=hein.journals/uflr75&i=167.

SAE International. J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, April 2021. URL https://www.sae.org/standards/content/j3016_202104/.

Salavati, C., Song, S., Diaz, W. S., Hale, S. A., Montenegro, R. E., Murai, F., and Dori-Hacohen, S. Reducing Biases towards Minoritized Populations in Medical Curricular Content via Artificial Intelligence for Fairer Health Outcomes. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1269–1280, October 2024. ISSN 3065-8365. doi: 10.1609/aies.v7i1.31722. URL https://ojs.aaai.org/index.php/AIES/article/view/31722. Number: 1.

Salem, A. H., Azzam, S. M., Emam, O. E., and Abohany, A. A. Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. *Journal of Big Data*, 11(1):105, August 2024. ISSN 2196-1115. doi: 10.1186/s40537-024-00957-y. URL https://doi.org/10.1186/s40537-024-00957-y.

Salvaggio, E. Most Researchers Do Not Believe AGI Is Imminent. Why Do Policymakers Act Otherwise? | TechPolicy.Press, March 2025. URL https://techpolicy.press/most-researchers-do-not-believe-agi-is-imminent-why-do-policymakers-act-otherwise.

Sartori, L. and Bocca, G. Minding the gap(s): public perceptions of AI and socio-technical imaginaries. *AI & SOCIETY*, 38(2):443–458, April 2023. ISSN 1435-5655. doi: 10.1007/s00146-022-01422-1. URL https://doi.org/10.1007/s00146-022-01422-1.

Saxon, M., Holtzman, A., West, P., Wang, W. Y., and Saphra, N. Benchmarks as microscopes: A call for model metrology. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=bttKwCZDkm.

Scheuerman, M. K., Hanna, A., and Denton, E. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW):1–37, 2021.

Schulz, A. J., Krieger, J., and Galea, S. Addressing Social Determinants of Health: Community-Based Participatory Approaches to Research and Practice. *Health Education & Behavior*, 29(3):287–295, June 2002. ISSN 1090-1981. doi: 10.1177/109019810202900302. URL https://doi.org/10.1177/109019810202900302. Publisher: SAGE Publications Inc.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., and Young, M. Machine learning: The high interest credit card of technical debt. In *SE4ML: software engineering for machine learning (NIPS 2014 Workshop)*, volume 111, pp. 112. Cambridge, MA, 2014.

Searle, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, September 1980. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00005756. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68, Atlanta GA USA, January 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287598. URL https://dl.acm.org/doi/10.1145/3287560.3287598.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, July 2022. doi: 10.1109/ijcnn55064.2022.9891914. URL http://dx.doi.org/10.1109/IJCNN55064.2022.9891914.

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, pp. 723–741, New York, NY, USA, August 2023. Association for Computing Machinery. ISBN 9798400702310. doi: 10.1145/3600211.3604673. URL https://doi.org/10.1145/3600211.3604673.

Shi, F. and Evans, J. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14(1):1641, March 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36741-4. URL https://www.nature.com/articles/s41467-023-36741-4. Publisher: Nature Publishing Group.

Shilton, K. Values and Ethics in Human-Computer Interaction. *Foundations and Trends® in Human–Computer Interaction*, 12(2):107–171, 2018. ISSN 1551-3955, 1551-3963. doi: 10.1561/1100000073. URL http://www.nowpublishers.com/article/Details/HCI-073.

Siler, K., Lee, K., and Bero, L. Measuring the effectiveness of scientific gatekeeping. *Proceedings of the National Academy of Sciences*, 112(2):360–365, 2015. doi: 10.1073/pnas.1418218112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1418218112.

Simonton, D. K. Psychology's Status as a Scientific Discipline: Its Empirical Placement within an Implicit Hierarchy of the Sciences. *Review of General Psychology*, 8(1):59–67, March 2004. ISSN 1089-2680. doi: 10.1037/1089-2680.8.1.59. URL https://doi.org/10.1037/1089-2680.8.1.59.

Sloane, M., Moss, E., and Chowdhury, R. A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns*, 3(2), February 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100425. URL https://www.cell.com/patterns/abstract/S2666-3899(21)00308-1. Publisher: Elsevier.

Smart, A. *Beyond zero and one: machines, psychedelics, and consciousness*. OR Books, New York, 2015. ISBN 978-1-68219-006-7.

Soderberg, C. K., Errington, T. M., and Nosek, B. A. Credibility of preprints: an interdisciplinary survey of researchers. *Royal Society Open Science*, 7(10): 201520, October 2020. doi: 10.1098/rsos.201520. URL https://royalsocietypublishing.org/doi/full/10.1098/rsos.201520. Publisher: Royal Society.

Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., and Choi, Y. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, March 2024a. ISSN 2374-3468. doi: 10.1609/aaai.v38i18.29970. URL https://ojs.aaai.org/index.php/AAAI/article/view/29970. Number: 18.

Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=gQpBnRHwxM.

Srivastava, T., Chou, J.-C., Shroff, P., Livescu, K., and Graziul, C. Speech Recognition For Analysis of Police Radio Communication. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 906–912, December 2024. doi: 10.1109/SLT61566.2024.10832157. URL https://ieeexplore.ieee.org/document/10832157/metrics#metrics.

Stirling, A. Disciplinary dilemma: working across research silos is harder than it looks. *The Guardian*, 11:1–4, 2014.

Stokols, D., Fuqua, J., Gress, J., Harvey, R., Phillips, K., Baezconde-Garbanati, L., Unger, J., Palmer, P., Clark, M. A., Colby, S. M., et al. Evaluating transdisciplinary science. *Nicotine & tobacco research*, 5(Suppl_1):S21–S39, 2003.

Suchman, L. The uncontroversial 'thingness' of AI. *Big Data & Society*, 10(2):20539517231206794, July 2023. ISSN 2053-9517. doi: 10.1177/20539517231206794. URL https://doi.org/10.1177/20539517231206794. Publisher: SAGE Publications Ltd.

Suleyman, M. and Bhaskar, M. *The Coming Wave*. Crown, New York, first edition edition, 2023. ISBN 978-0-593-59396-7.

Summerfield, C. *Natural general intelligence: how understanding the brain can help us build AI*. Oxford University Press, Oxford New York, NY, first edition edition, 2023. ISBN 978-0-19-284388-3.

Tenopir, C., Levine, K., Allard, S., Christian, L., Volentine, R., Boehm, R., Nichols, F., Nicholas, D., Jamali, H. R., Herman, E., and Watkinson, A. Trustworthiness and authority of scholarly information in a digital age: Results of an international questionnaire. *Journal of the Association for Information Science and Technology*, 67(10): 2344–2361, 2016. ISSN 2330-1643. doi: 10.1002/asi.23598. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23598.

The Royal Society. *Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research*. The Royal Society, United Kingdom, May 2024.

Tibebu, H. DeepSeek and the Race to AGI: How Global AI Competition Puts Ethical Accountability at Risk | TechPolicy.Press. *Tech Policy Press*, January 2025. URL https://techpolicy.press/deepseek-and-the-race-to-agi-how-global-ai-competition-puts-ethical-accountability-at-risk.

United Nations. Governing AI for Humanity. Final Report, United Nations, New York, NY, September 2024. URL https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

Van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., and Rich, P. Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, pp. 1–21, 2024.

Veit, W. Model Pluralism. *Philosophy of the Social Sciences*, 50(2):91–114, March 2020. ISSN 0048-3931, 1552-7441. doi: 10.1177/0048393119894897. URL http://journals.sagepub.com/doi/10.1177/0048393119894897.

Venkit, P. N., Graziul, C., Goodman, M. A., Kenny, S. N., and Wilson, S. Race and Privacy in Broadcast Police Communications. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2):382:1–382:26, November 2024. doi: 10.1145/3686921. URL https://dl.acm.org/doi/10.1145/3686921.

Vestal, A. and Mesmer-Magnus, J. Interdisciplinarity and team innovation: The role of team experiential and relational resources. *Small Group Research*, 51(6):738–775, 2020.

Viljoen, S. A Relational Theory of Data Governance. *The Yale Law Journal*, 2021.

Wang, A., Kapoor, S., Barocas, S., and Narayanan, A. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM Journal on Responsible Computing*, 1(1):1–45, March 2024. ISSN 2832-0565. doi: 10.1145/3636509. URL https://dl.acm.org/doi/10.1145/3636509.

Wang, D., Prabhat, S., and Sambasivan, N. Whose AI Dream? In search of the aspiration in data annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pp. 1–16, New York, NY, USA, April 2022. Association for Computing Machinery. ISBN 978-1-4503-9157-3. doi: 10.1145/3491102.3502121. URL https://dl.acm.org/doi/10.1145/3491102.3502121.

Warne, R. T. and Burningham, C. Spearman's g found in 31 non-Western nations: Strong evidence that g is a universal phenomenon. *Psychological Bulletin*, 145(3):237–272, March 2019. ISSN 0033-2909. doi: 10.1037/bul0000184. URL http://proxy.uchicago.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=2019-01683-001&site=ehost-live&scope=site.

Weidinger, L., Mellor, J. F. J., Pegueroles, B. G., Marchal, N., Kumar, R., Lum, K., Akbulut, C., Diaz, M., Bergman, A. S., Rodriguez, M. D., Rieser, V., and Isaac, W. STAR: SocioTechnical Approach to Red Teaming Language Models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21516–21532, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1200. URL https://aclanthology.org/2024.emnlp-main.1200/.

Weizenbaum, J. *Computer power and human reason: from judgment to calculation*. Freeman, San Francisco, 1976. ISBN 978-0-7167-0464-5 978-0-7167-0463-8.

Whitney, C. D. and Norman, J. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1733–1744, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659002. URL https://dl.acm.org/doi/10.1145/3630106.3659002.

Widder, D. G. Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 1295–1304, New York, NY, USA, June 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658973. URL https://dl.acm.org/doi/10.1145/3630106.3658973.

Widder, D. G. and Hicks, M. Watching the Generative AI Hype Bubble Deflate, August 2024. URL http://arxiv.org/abs/2408.08778. arXiv:2408.08778.

Widder, D. G. and Nafus, D. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1): 20539517231177620, January 2023. ISSN 2053-9517. doi: 10.1177/20539517231177620. URL https://doi.org/10.1177/20539517231177620. Publisher: SAGE Publications Ltd.

Xu, F., Wu, L., and Evans, J. Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences*, 119(23):e2200927119, June 2022. doi: 10.1073/pnas.2200927119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2200927119. Publisher: Proceedings of the National Academy of Sciences.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. Benchmarking benchmark leakage in large language models, 2024. URL https://arxiv.org/abs/2404.18824.

Young, M., Rodriguez, L., Keller, E., Sun, F., Sa, B., Whittington, J., and Howe, B. Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 191–200, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287577. URL https://dl.acm.org/doi/10.1145/3287560.3287577.

Young, M., Ehsan, U., Singh, R., Tafesse, E., Gilman, M., Harrington, C., and Metcalf, J. Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday*, April 2024. ISSN 1396-0466. doi: 20240428092301000. URL https://firstmonday.org/ojs/index.php/fm/article/view/13642.

Yu, D., Rosenfeld, H., and Gupta, A. The 'AI divide' between the Global North and Global South, January 2023. URL https://www.weforum.org/stories/2023/01/davos23-ai-divide-global-north-global-south/.

Zeff, M. Microsoft and OpenAI have a financial definition of AGI: Report, 2025. URL https://techcrunch.com/2024/12/26/microsoft-and-openai-have-a-financial-definition-of-agi-report/. [Online; accessed 17-January-2025].

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Zhuang, C., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., and Yue, S. A Careful Examination of Large Language Model Performance on Grade School Arithmetic. *Advances in Neural Information Processing Systems*, 37:46819–46836, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/53384f2090c6a5cac952c598fd67992f-Abstract-Datasets_and_Benchmarks_Track.html.

Zhang, L., Sun, B., Jiang, L., and Huang, Y. On the relationship between interdisciplinarity and impact: Distinct effects on academic and broader impact. *Research Evaluation*, 30(3):256–268, July 2021. ISSN 0958-2029. doi: 10.1093/reseval/rvab007. URL https://doi.org/10.1093/reseval/rvab007.

Zhao, D., Andrews, J., Papakyriakopoulos, O., and Xiang, A. Position: Measure dataset diversity, don't just claim it. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60644–60673. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhao24a.html.

Zhu, Z. Paradigm, specialty, pragmatism: Kuhn's legacy to methodological pluralism. *Systems Research and Behavioral Science*, 39(5):895–912, 2022. ISSN 1099-1743. doi: 10.1002/sres.2881. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sres.2881. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sres.2881.

A. Definitions of AGI and Related Concepts

Table 1 below presents illustrative definitions of AGI and usefully related concepts. We agree with Morris et al. (2024)'s proposal to broaden discussions of AGI definitions to include accounts that avoid the term "AGI" yet address similar goals of achieving human-level intelligence. For example, while OpenAI's influential definition (OpenAI, 2018) focuses on outperforming humans at economically valuable work, sharing key parallels with Nilsson (2005). Yet it notably differs from Chollet et al. (2024); Summerfield (2023); Morris et al. (2024); Chollet (2019); Goertzel (2014) and others by not explicitly emphasizing generality.

Following Blili-Hamelin et al. (2024), we believe discussions of AGI definition should include approaches that challenge AGI's central premises. Below, we include Weizenbaum (1976) and Attard-Frost (2023). Including these critical accounts enables noticing a surprising similarity with Summerfield (2023)'s reconceptualization of AGI through the lens of natural intelligence: all three accounts favor a strong form of contextualism and pluralism about what intelligence means.

B. AGI as a North-Star Goal

This paper argues against AGI serving as a north-star goal of AI research. When we talk about AGI being treated as north-star goal, we are not claiming that the majority of AI researchers are explicitly working in pursuit of this goal. In fact, many AI researchers may, like us, doubt or reject this goal (Salvaggio, 2025; Francesca Rossi et al., 2025). Rather, we claim that influential researchers and executives hold this view, enough so for it to deserve scrutiny. These dominant voices are further amplified by the publicity that discussions of AGI generate, including by members of the press (Klein, 2025), and by government commissions (Bommasani et al., 2025). As such, AGI has come to permeate both community incentives and cultural norms. In this context, interrogating its role and influence in the AI research community matters. Below we provide a small sample of quotes and resources illustrating this effect.

OpenAI The mission statement of OpenAI is "…to ensure that [AGI] benefits all of humanity." Its website (OpenAI, 2025a) states that "we are building safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome." This goal directly influences both the company's direct work (OpenAI, 2025b) and the work that it funds (OpenAI, 2025c).

Google DeepMind The vision statement (DeepMind, 2025) of Google DeepMind states that "[AGI] has the potential to drive one of the greatest transformations in history." In a recent briefing (Browne, 2025), Demis Hassabis stated that, though current systems still have limitations, over the next 5–10 years "a lot of those capabilities will start coming to the fore and we'll start moving towards what we call [AGI]." A position paper (Morris et al., 2024) by Google DeepMind authors last year defines concrete goals in pursuit of AGI.

Anthropic In the essay "Machines of Loving Grace," Anthropic CEO Dario Amodei argues how the world could be shaped positively by "Powerful AI" aligned with "AGI" goals (Amodei, 2024). Amodei recently told CNBC that AI that is "better than almost all humans at almost all tasks" can emerge shortly (Browne, 2025). Anthropic's official recommendations (Anthropic, 2025) to OSTP for the U.S. AI Action Plan state that "we expect powerful AI systems will emerge [with] intellectual capabilities matching or exceeding that of Nobel Prize winners."

Other Influential Executives & Researchers In "Sparks of AGI" (Bubeck et al., 2023), researchers at Microsoft argue that GPT-4 "could reasonably be viewed as an early (yet still incomplete) version of [AGI]." In announcing xAI, Elon Musk stated (Kerner, 2023) that "the overarching goal of xAI is to build a good AGI." Speaking to TIME (Perrigo, 2024), Yann LeCunn explained that he refers to "what people call 'AGI'" as "human-level intelligence," and noted that "the mission of FAIR [Meta's Fundamental AI Research team] is human-level intelligence." Geoff Hinton stated to Forbes (Maymin, 2023) that he "is certain we will have AGI soon, and biological humans will be relegated to be the second-smartest species on the planet."

Table 1: Sample of proposed definitions of AGI and related concepts.

Weizenbaum (1976). "Intelligence is a meaningless concept in and of itself. It requires a frame of reference, a specification of a domain of thought and action, in order to make it meaningful. [. . .] [T]hese domains are themselves not measurable." Argues that any argument that calls for the conclusion or denial that "machines may surpass us in general intelligence" is "ill-framed and therefore sterile" due to "our inability to compute an upper bound on machine intelligence." We follow Blili-Hamelin et al. (2024) in considering this critical account relevant to debates about how to conceive AGI.
Searle (1980). "according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states."
Gubrud (1997) "By advanced artificial general intelligence, I mean AI systems that rival or surpass the human brain in complexity and speed, that can acquire, manipulate and reason with general knowledge, and that are usable in essentially any phase of industrial or military operations where a human intelligence would otherwise be needed. Such systems may be modeled on the human brain, but they do not necessarily have to be, and they do not have to be "conscious" or possess any other competence that is not strictly relevant to their application. What matters is that such systems can be used to replace human brains in tasks ranging from organizing and running a mine or a factory to piloting an airplane, analyzing intelligence data or planning a battle."
Nilsson (2005). "achieving real human-level artificial intelligence would necessarily imply that most of the tasks that humans perform for pay could be automated. Rather than work toward this goal of automation by building special-purpose systems, I argue for the development of general-purpose, educable systems that can learn and be taught to perform any of the thousands of jobs that humans can perform."
Fast Company (2010) "*Wozniak: Could a Computer Make a Cup of Coffee?*" tasks the machine to go into an "average" American home, find ingredients, and make a cup of coffee. This requires embodied AI systems. Wozniak's test has since been included in discussions of AGI (Goertzel et al., 2012).
Chalmers (2010). "AI is artificial intelligence of human level or greater (that is, at least as intelligent as an average human). Let us say that AI+ is artificial intelligence of greater than human level (that is, more intelligent than the most intelligent human). Let us say that AI++ (or superintelligence) is AI of far greater than human level (say, at least as far beyond the most intelligent human as the most intelligent human is beyond a mouse)."
Goertzel et al. (2012). Propose an architecture for human-like general intelligence that integrates slightly modified versions of previously existing architectures, emphasizing the commonalities across different approaches.
Bostrom (2014). "We can tentatively define a superintelligence as *any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*."
Goertzel (2014). "roughly speaking, an AGI system is a synthetic intelligence that has a general scope and is good at generalization across various goals and contexts."
Smart (2015). "a strong AI system would be an entirely autonomous computer system in no way controlled or influenced by human operators. It could successfully adapt to its environment or even be part of its environment, making intelligent decisions, and for all intents and purposes interacting with humans naturally. It would have vastly superior memory and computational abilities but would also be able to reason and act accordingly. What all of this boils down to is that a strong AI would have to be conscious."
OpenAI (2018). "OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity." December 2024 reporting suggests that OpenAI and Microsoft "signed an agreement last year stating OpenAI has only achieved AGI when it develops AI systems that can generate at least \$100 billion in profits" (Zeff, 2025). If true, this is a significant departure from their former definition.
Chollet (2019); Chollet et al. (2024). Defines AGI as "a system capable of efficiently acquiring new skills and solving novel problems for which it was neither explicitly designed nor trained." In 2019, introduced an as yet (January 2025) unsolved benchmark for incentivizing progress towards AGI thus defined. Proposes that "it's still feasible to create unsaturated, interesting benchmarks that are easy for humans, yet impossible for AI – without involving specialist knowledge. We will have AGI when creating such evals becomes outright impossible" (Chollet, 2024b).
Hernández-Orallo et al. (2021). "independently of its overall capability, *an agent can only be called fully general if it covers all tasks up to an equivalent level of difficulty, determined by the resources that are needed for them*." Introduces two individual-specific (be them humans, other animals or AI systems) measures that, together, decouple the concept of general intelligence: (i) 'generality', which refers to "the *distribution* of the tasks the agent can solve," and (ii) 'capability', which indicates "how far, on average, an agent can reach in terms of task difficulty."
Marcus (2022). Defines AGI as "a shorthand for any intelligence. . . that is flexible and general, with resourcefulness and reliability comparable to (or beyond) human intelligence." Calls for the need to operationalize the definition as a single system that can succeed in at least 3 of 5 proposed tasks, including Wozniak's coffee cup benchmark (Fast Company, 2010).

Morris et al. (2024). Proposes a practical strategy analogous to Levels of Driving Automation standards (SAE International, 2021). Describes a graded set of levels of achievement of target characteristics that can each be associated with tangible "metrics", the introduction of "risks", and changes in "Human-AI Interaction paradigm" (Morris et al., 2024). The framework targets 2 characteristics: levels of performance (which they define as "the depth of an AI system's capabilities, i.e., how it compares to human-level performance for a given task"), and levels of generality (defined as "breadth of an AI system's capabilities, i.e., the range of tasks for which an AI system reaches a target performance threshold").
Summerfield (2023) We consider this account to endorse a strong form of pluralism about intelligence: natural intelligence takes many different shapes across cultures and species, serving many goals and functions that are irreducibly shaped by "the internal model by which an animal understands the world", which itself "depends on its local environment, its embodied form, its desires and goals, and its interactions with conspecifics." The same should be expected for "strong AI" or "AGI". However, building on Dreyfus & Dreyfus (1986), Summerfield argues we have practical reasons to constrain the forms AI takes. The goal of AI is "to help humans in their endeavours." To that end,"if we want to build AI systems that exhibit human-like intelligence, with whom we can interact in pursuit of human-centred goals, these agents will need to think in ways that make sense to us."[10]
Attard-Frost (2023) Defines human and artificial intelligence as "value-dependent cognitive performance", and "centres interdependencies between agents, their environments, and their measurers in collectively constructing and measuring context-specific performances of intelligent action." Although this account is not presented as a conception of AGI, Blili-Hamelin et al. (2024) argue that the account is relevant to the topic.
Suleyman & Bhaskar (2023) "Artificial intelligence (AI) is the science of teaching machines to learn humanlike capabilities. Artificial general intelligence (AGI) is the point at which an AI can perform all human cognitive skills better than the smartest humans."
Agüera y Arcas & Norvig (2023). "'General intelligence' must be thought of in terms of a multidimensional scorecard, not a single yes/no proposition." Dimensions discussed include topics, tasks, modalities, languages, and instructability.

[10] "[W]e are building AI to make the world a better place. But if we want AI to be useful to people, it will need to share our umwelt. If we build an AI that sees the world in a radically different way to us, its behaviour and mental states will be unintelligible. Such an agent will be at best unreliable and at worst unsafe." (Summerfield, 2023)

Author Contributions

We follow the CRediT recommendations and taxonomy provided by Allen et al. (2019) to determine and outline author contributions.[11]

- Borhane Blili-Hamelin: Conceptualization (Formulation & Evolution), Investigation, Methodology (Development), Project administration, Supervision (Oversight & Leadership), Writing (Initial draft, Submitted draft, Review & Editing).
- Christopher Graziul: Conceptualization (Formulation & Evolution), Investigation, Methodology (Development), Project administration, Supervision (Oversight & Leadership), Writing (Initial draft, Submitted draft, Review & Editing).
- Leif Hancox-Li: Conceptualization (Formulation & Ideas), Investigation, Methodology (Development), Supervision (Mentorship), Writing (Initial draft, Submitted draft, Review & Editing).
- Hananel Hazan: Conceptualization (Ideas & Evolution), Investigation, Methodology (Development), Project administration, Writing (Initial draft, Review & Editing, LaTeX).
- El-Mahdi El-Mhamdi: Conceptualization (Evolution & Ideas), Writing (Initial draft, Review & Editing).
- Avijit Ghosh: Conceptualization (Ideas), Writing (Submitted draft [Traps section], Review & Editing).
- Katherine Heller: Conceptualization (Formulation & Evolution), Supervision (Mentorship), Writing (Initial draft, Submitted draft, Review & Editing).
- Jacob Metcalf: Conceptualization (Formulation & Evolution), Writing (Submitted draft, Review & Editing).
- Fabricio Murai: Conceptualization (Ideas), Writing (Submitted draft [Table of AGI definitions], Review & Editing, LaTeX).
- Eryk Salvaggio: Conceptualization (Evolution & Ideas), Writing (Submitted draft [Introduction, Traps section], Review & Editing).
- Andrew Smart: Conceptualization (Formulation & Evolution), Writing (Initial draft, Review & Editing).
- Todd Snider: Conceptualization (Ideas), Methodology (Implementation), Writing (Initial draft, Review & Editing, LaTeX).
- Mariame Tighanimine: Conceptualization (Evolution & Ideas), Writing (Initial draft, Review & Editing).
- Talia Ringer: Conceptualization (Formulation, & Ideas), Project administration, Supervision (Oversight & Leadership), Writing (Initial draft, Review & Editing).
- Margaret Mitchell: Conceptualization (Ideas & Evolution), Investigation, Methodology (Development), Supervision (Oversight & Mentorship), Writing (Initial draft, Submitted draft [all sections], Review & Editing, Rebuttal).
- Shiri Dori-Hacohen: Conceptualization (Ideas & Evolution), Investigation, Methodology, Project administration, Supervision (Oversight & Leadership), Writing (Initial draft, Submitted draft [all sections], Review & Editing).

[11]The initial ICML submission was created by Borhane Blili-Hamelin, Leif Hancox-Li, and Christopher Graziul, leveraging extensive work and writing from the larger project. All contributors then worked together on refining this submission.