

Whose “Intelligence” Are We Building? On Exclusion in Conceptual Frameworks of Artificial Intelligence

Dafna Burema*, Vito Mengers*

Technische Universität Berlin, Germany
Science of Intelligence, Research Cluster of Excellence, Berlin, Germany
burema@tu-berlin.de, v.mengers@tu-berlin.de

Abstract

Every definition of intelligence differentiates between haves and have nots, yet the exclusionary premises embedded in different conceptual frameworks of intelligence are rarely examined. This paper argues these premises matter both normatively and practically, as they are value-laden and operationalized in AI benchmarks used to evaluate “intelligent” systems. Drawing on Science and Technology Studies (STS) and intelligence research, we examine three dominant conceptual frameworks, i.e., anthropocentrism, pure ability, and generality, and show how they entail distinct speciesist and ableist assumptions that partially manifest in current AI evaluation practices. Rather than advocating for a universally inclusive definition of intelligence, we argue for epistemic reflexivity. To support this, we propose a structured set of questions for AI developers to make the normative assumptions embedded in benchmark design explicit.

1 Introduction

Developing Artificial Intelligence requires an interpretation of “intelligence.” The field of intelligence research does not have a single agreed-upon definition and is currently performing multiple conceptualizations in different strands of research. What is consistent and inherent to the definition is an exclusionary element: the concept relies on haves and have nots. After all, claiming that something, someone, or some behavior “is” intelligent relies on the notion that something, someone, or some behavior is *not* intelligent. The question then becomes who or what gets excluded with what definition of intelligence. The implications of labeling something or a behavior as more or less intelligent can however become problematic, since intelligence is a so-called “value-laden” concept (Cave 2020; Blili-Hamelin, Hancox-Li, and Smart 2024), and is therefore not a purely descriptive exercise. Intelligence is seen as something that is beneficial to have, develop or exhibit, and hence carries a normative element. This paper argues that it is therefore necessary to understand the exclusionary premises behind different interpretations of this concept. Furthermore, this paper argues that these exclusions could inform further technology development, such as AI benchmarks, and thereby draws on

scholarship in Science and Technology Studies (STS).

By doing so, this paper builds on the idea that defining intelligence is seen as a social process and therefore its definition a social construct (similar to Fischbach, Vandemeulebroucke, and Van Wynsberghe 2023), and that engineering is a normative practice (Pinch and Bijker 1984; Bijker 2010; Verbeek 2006). Rather than proposing a new definition of intelligence or judging existing frameworks and AI benchmarks as correct or incorrect, this paper argues for epistemic reflexivity in AI development and intelligence research communities. Different conceptualizations of intelligence carry distinct forms of epistemic exclusion, making it important to raise awareness of how ideas about who or what “is” intelligent relate to engineering practices such as benchmarks in AI. This paper therefore offers an overview of who or what is excluded with which conceptualization, and how these exclusions shape AI evaluation practices.

To elaborate, developing and deploying technoscience does not occur in a vacuum and needs to be understood in their broader societal meaning, which includes power asymmetries, exclusion, bias and discrimination. Several authors have traced the rather exclusionary and problematic historical background of intelligence research. For instance, Fischbach, Vandemeulebroucke, and Van Wynsberghe (2023) mention how the Turing test is related to colonialism, Cave (2020) more broadly gives an historic overview of the misuse of the concept of intelligence, Gebru and Torres (2024) discuss how different actors aiming to develop AGI have problematic viewpoints rooted in racism and eugenics, and also Shevlin et al. (2019) mention how historically intelligence is used as a classification tool for different marginalized groups. While there is a consensus that one should be careful with classifying and defining who or what gets to be “intelligent,” intelligence research operationalizes different definitions of intelligence, classifying different species and behaviors as “intelligent.” The question then becomes, who or what gets excluded with which conceptual framework, and what does it mean for AI development and technology construction?¹ In this paper, we address this gap in the liter-

¹Blili-Hamelin, Hancox-Li, and Smart (2024) are one of few authors who acknowledge the relevance of studying the different definitions of intelligence and their normativities, yet in their paper focus on AGI specifically and conflate definitions with visions, discourse, and benchmarks/test metrics.

*These authors contributed equally.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ature, by discussing 1) how *different conceptual frameworks* of intelligence carry *different exclusionary meanings*, and 2) how to foster reflexive epistemic engagement when constructing AI evaluation benchmarks. Concerning the former we trace the genealogy of the concept and argue that our understanding of intelligence over the years has rested on three pillars, i.e., anthropocentric intelligence, pure ability, and generality, and that in addition to the connotation and misuse by different actors of the concept in eugenics, racism, and sexism that prior literature states (e.g., Cave 2020; Gebru and Torres 2024), speciesist and ableist assumptions are also embedded at a conceptual level and in AI benchmarks. Concerning the latter, we propose a set of questions designed to foster epistemic reflexivity. Overall, this paper aims to raise awareness within the AI engineering community about the normative element to engineering intelligence, and foster reflexive practices for designing and choosing benchmarks.

This paper starts by discussing theories in STS in Section 2 that argue that technoscience is socially constructed, and the role of reflexivity in AI development. Thereafter, in Section 3, the paper outlines three major conceptual frameworks of intelligence, followed by Section 4, that describes how these conceptualizations are operationalized in AI benchmarks. Section 5 explains how these three conceptual frameworks of intelligence are related to speciesism and ableism. Finally, the paper discusses in Section 6 how to foster reflexivity.

2 Theoretical Framework

2.1 The Socially Constructed Nature of Technologies and Intelligence

In STS, many theories acknowledge the socially constructed nature of technologies. For instance, SCOT (social construction of technology) is a theory that shows how social processes shape the overall design of the artifact (Pinch and Bijker 1984; Bijker 2010). Machines are malleable, as different stakeholders get involved in the design process, ultimately influencing the overall outcome of the artifact. It therefore discards the idea of technoscience being a linear or inevitable process, but a social one. Similarly, other theories in STS also underscore the importance of understanding how social processes influence the construction of new technologies. For instance, inscription describes how technologies are developed with certain ideas inscribed in them, i.e., with certain instructions designed into the artifact to nudge users (Akrich 1992). However, this inscription is not without value or normativity (Verbeek 2006). For instance, a coffee cup has different affordances than a plastic cup: one is reusable and made from more sustainable materials than the other. At face value, AI systems may seem too general to analyze in terms of normative inscriptions, since situational context is lacking. To elaborate, the normative inscriptions differ per machine: AI in predictive policing has different normative ideas inscribed in them than elderly care robots. The former inscribes ideas about, e.g., police work, potential offenders, and neighborhoods, whereas the latter inscribes ideas about, e.g., caregiving and the welfare state. In other words, different machines are designed with different nor-

mativities in mind, making it difficult to speak of a single, unified set of normative inscriptions for “AI” as a whole.

Although a case-by-case analysis of AI’s inscription of particular worldviews is not feasible, it is possible to discern what normative ideas underlie the construction of “intelligence” as a concept. As prior literature argues, the central focus of “intelligence” in AI is not free of values and norms: “AI in particular is likely to reproduce a certain worldview because it explicitly imitates certain kinds of behavior which are based on a value-laden concept: “intelligence.” The basis of this concept originates from dominant societies of the Global North, reflected in the usage, aims, and design of the currently developed AI technologies.” (Fischbach, Vandemeulebroucke, and Van Wynsberghe 2023, p. 995). This quote does not only illustrate the need to understand the values inscribed into AI and what intelligence means, but also, the emphasis of the role of the Global North as the norm.

The focus on the Global North in the discussion on intelligence, illustrates how the concept of intelligence is situated in a wider discussion on post-colonialism. It is furthermore discussed in relation to classism, sexism, racism, and eugenics (Gebru and Torres 2024; Fischbach, Vandemeulebroucke, and Van Wynsberghe 2023; Cave 2020; Shevlin et al. 2019). To make these arguments, critical scholars analyzed either the actors behind such movements of (general) artificial intelligence and their obsession with “solving” intelligence (e.g., Cave 2020; Gebru and Torres 2024), or specific benchmarks and methods to assess intelligence (e.g., Fischbach, Vandemeulebroucke, and Van Wynsberghe 2023; Alfano, Holden, and Conway 2017).

Concerning actors, Cave (2020) argues that the developers of AI and the workforce of Silicon Valley are dominated by white men, who are using the concept “intelligence” to legitimize their position. His paper starts by describing how the definition of intelligence has changed over time, while upholding a certain normative hierarchy. For instance, during the 17th century, intellectual superiority was a way to justify colonialism and mastery over nature. The concept of intelligence creates legitimacy over who is capable of ruling. The advancements of science and technology, and intellectual merit was seen as a sign of superiority, while “inhabiting the opposite of this civilized state were those peoples deemed inferior in intellect and correspondingly pre-technological, and who were therefore savages” (Cave 2020, p. 30). Cave (2020) then makes a jump to current times and argues how the word “intelligence” according to him is “fetishized” in the AI community, and still implies a normative hierarchy with white men dominating the workforce in Silicon Valley. The paper focuses on the industry itself: *people* making the AI, their visions, hopes and fears. Similarly, Gebru and Torres (2024) aim to understand the normativities in building Artificial General Intelligence (AGI) by focusing on “primary sources by leading figures investing in, advocating for, and attempting to build AGI” (Gebru and Torres 2024, p. 2). In other words, similar to Cave (2020), their analysis on normativity focuses on actors and their visions. In this paper, however, the focus is not on the actors and their visions on AI or AGI, rather, how different research strands and their conceptual frameworks carry normativities and exclusion.

In addition to focusing on actors in the field, previous literature focuses on biased benchmarks and other tools to assess intelligence. To give an example, Alfano, Holden, and Conway (2017) give a historical overview of intelligence testing. One of the first tests to measure intelligence was developed to identify learning disabilities, but it grew over the years as a means to study populations and to be used for institutional admissions. The paper mentions several critiques for these variations of intelligence tests: they are typically too narrow in the language used for testing, and are not inclusive of language that is spoken in certain communities. Alfano, Holden, and Conway (2017) mention the word “regatta,” as a word used in one of the questions asked in the US college admission test SAT, but this word reflects a certain cultural capital that is not broadly shared by the general population, but the ruling elite. In other words, there are no intelligence tests that are free of cultural influences. This is similar to the argument Fischbach, Vandemeulebroucke, and Van Wynsberghe (2023) are making about the Turing Test: It reflects the global north’s conception of intelligence according to the authors. Different measures of intelligence could reflect different normative biases. The underlying notion rests on a hierarchy of knowledge: some forms of knowledge typically associated with certain groups having more epistemic capital than others. The concept of “intelligence” by definition makes a distinction between haves and have nots of intelligence. This mere labeling in itself is not an act of injustice, but since it is a value-laden and normative concept it needs some careful consideration as to why certain individuals, groups, or behaviors are seen as (not) intelligent. If certain individuals, groups, or behaviors are not included in the conceptualization because of biased or narrow standards, it would become ethically problematic. To mitigate this, AI engineers and intelligence researchers should be aware of the categorical exclusion they are choosing and engineering in their work. This could be fostered with practices of reflexivity as is explained next.

2.2 Reflexivity

In qualitative research, scholars are trained to critically reflect on their positionality while conducting fieldwork (Lindlof and Taylor 2017; Burema 2025). This entails, to be aware of one’s personal identity and biases that could potentially influence the study findings and outcomes. It relies on the idea that objectivity is not something that can be extracted as such, since scholars themselves have lived experiences and identities that they bring into their research, and simultaneously influence the field they are observing. Reflexivity is an ongoing process, in which researchers question how their beliefs shape what they observe and study.

The field of intelligence research, computer science and AI engineering do not encourage reflecting on the social identity of their scholars per se, as STEM research does not necessarily engage in anthropological or ethnographic fieldwork. However, in the context of AI ethics, reflexivity is implicitly applied in the principles of fairness and transparency. To elaborate, methods of documentation are proposed and adopted in the machine learning community to critically reflect where data comes from, and how models

are constructed. Reasons for doing so, stem from bias mitigation, as datasheets for datasets (Geburu et al. 2021) could show how a particular dataset is annotated and how it (does not) adequately represent certain parts of a population. This process includes a reflection on the process of data acquisition: how was it made with whom in mind? Who decides what is ethically acceptable or not? These questions are part and parcel of reflexive practices. It makes taken-for-granted assumptions about engineering practices explicit, allowing for critical scrutiny, albeit without reflecting on the social position of the researcher themselves.

Famously, model cards for model reports (Mitchell et al. 2019) and datasheets for datasets (Geburu et al. 2021) are established practices of documentation and reflexivity in the field of AI ethics. Specifically, for benchmarks in AI, similar methods for reflexivity are proposed such as Eval Factsheets (Bordes et al. 2025) and BenchmarkCards (Sokol et al. 2025). The latter example also includes questions on ethics, yet, these are related procedural ethics such as compliance checks and consent procedures. This present paper takes a different angle on ethics, as it asks AI developers to critically reflect how “intelligence” is ontologically inscribed into benchmarks. Section 6 discusses a set of questions that foster this type of reflexivity.

3 Conceptual Frameworks of Intelligence

The term “intelligence” has accumulated a wide array of definitions over time (Legg and Hutter 2007), with roots in philosophy, psychology, and computer science (Peterson 1925; Gonzalez 1979; Mackintosh 2011). These interpretations differ not only in the mechanisms they propose, e.g. symbolic manipulation of formal structures (McCarthy et al. 1955; Simon and Newell 1971), neural computation and learning from data (Rosenblatt 1958; Berkeley 2019), morphological computation (Pfeifer and Gómez 2009), or computation more broadly (Brock 2024), but also in their underlying assumptions about what counts as intelligence in the first place. Embodied and morphological views remain in the minority within mainstream AI and cognitive science, though the locus of intelligence has shifted from the brain toward the embodied agent in its environment (Varela, Thompson, and Rosch 1992), and explanations still differ on whether intelligence is primarily innate or learned (Bouchard Jr 1998; Polderman et al. 2015).

Rather than delve into the mechanisms of intelligence, we focus on the different conceptual frameworks that guide how intelligence is modeled, evaluated, and pursued in both natural and artificial domains. These include *anthropocentrism* (Section 3.1), *pure ability* (Section 3.2), and *generality* (Section 3.3). They offer different reference points for evaluating intelligence and correspond to long-standing tensions in AI between modeling human cognition, achieving goal-directed performance, and building systems that can function flexibly across diverse settings. Although these orientations are not mutually exclusive, they often focus on different goals in AI research and development, and they continue to influence both academic discourse and public understanding of intelligence.

A note on scope is warranted here. We do not claim that these three frameworks exhaust the ways intelligence can be conceptualized, nor that they are mutually exclusive in practice. They are dominant rather than complete: most benchmarks invoke more than one, and a single benchmark may operationalize generality in its stated aims while relying on anthropocentric reference points in its construction. Furthermore, other axes of exclusion, such as cultural capital and language, do not constitute a fourth framework but cut *across* all three. The “regatta” example discussed in Section 2 illustrates this: a test may be framed entirely in terms of ability while nonetheless including classist bias. These exclusions are not visible from the taxonomy alone, yet can still manifest in AI benchmarks. Therefore, the questions we propose in Section 6 give space to reflect on such cross-cutting exclusions.

3.1 Anthropocentrism – As Human as Possible

One enduring perspective of intelligence equates it with human cognition and behavior. This anthropocentric orientation can be divided into two interpretations: acting humanly, where intelligence is judged based on behavior indistinguishable from that of a human, and thinking humanly, where the focus is on replicating internal cognitive mechanisms assumed to be human (Russell, Norvig, and Davis 2010). Another, more common version treats humans as the default reference standard: other systems are considered intelligent to the extent they approximate human-relevant traits such as language use, problem-solving, or self-awareness (Shettleworth 2009; De Waal 2016).

Historically, intelligence was often framed as the distinguishing trait of humans, separating us from animals. This assumption underpinned classical philosophy—from Aristotle to British Empiricism—and carried into early cognitive science and AI (Peterson 1925). In early AI research, this led to both behaviorist approaches, exemplified by the Turing Test (Turing 1950, pp. 446–447), and cognitive models aiming to replicate human reasoning processes (McCarthy et al. 1955). While modern AI has largely shifted toward other benchmarks (as reflected in current definitions (Legg and Hutter 2007; Chollet 2019)), anthropocentrism persists in public discourse and technical practice.

Foundational textbooks in AI still present human-likeness as one plausible framing of intelligence (Russell, Norvig, and Davis 2010). In fields like Human-Computer Interaction, perceived intelligence is often tied to how closely machines resemble human communication or expression. For instance, large language models are frequently regarded as “intelligent” due to their fluency and responsiveness in human-like conversation, even when their reasoning capabilities are limited (Bender et al. 2021; Mei et al. 2024).

In AI development, this orientation is embodied in methods like imitation learning, where agents are trained to replicate human behaviors from demonstrations (Billard, Calinon, and Dillmann 2016; Chi et al. 2025; Zare et al. 2024). Mimicking human behavior in such systems is often intended not to define or replicate intelligence per se, but to improve human-machine interaction or leverage existing human expertise. While imitation is just one of many ap-

proaches to building intelligent systems, its prevalence can reinforce a model of intelligence that centers human norms and behaviors (Rajaraman et al. 2020).

3.2 Pure Ability – As “Able” as Possible

A second perspective conceptualizes intelligence as the capacity to perform “well”, which is defined in terms of goal-directed success, decision quality, or task competence. This perspective aligns with rational-agent models in AI, where intelligence is understood as the power to achieve objectives efficiently, regardless of how human-like the behavior appears (Russell, Norvig, and Davis 2010; Russell 1997).

Here, “rationality” refers not to prototypically human thought, but to the notion of maximizing expected utility under constraints. This view is central to different AI methods, such as planning, control, and reinforcement learning, and it often merges with a more general concern for “ability” in the sense of task performance. While “rationality” and “ability” are distinct concepts, i.e., rationality concerns decision optimality, ability concerns capacity to succeed, they often converge in practice through performance metrics.

Historically, the ability perspective grew in prominence during efforts to measure and compare human capacities, most notably in the rise of psychometrics and IQ testing in the late 19th and early 20th centuries (Gonzalez 1979; Peterson 1925). While early psychologists like Wundt focused on common capacities across humans, others emphasized inter-individual differences and designed tests to rank individuals by specific abilities. Over time, this view fragmented into multiple “intelligences,” such as logical, musical, emotional, and social (Gardner and Hatch 1989; Goleman 1995), all of which define intelligence by domain-specific success. To have high musical intelligence for instance, one must have sensitivity to pitch, meter, melody, and timbre (Gardner and Hatch 1989); for high emotional or interpersonal intelligence, one must show awareness of others’ feelings, temperaments, and motivations, along with strong communication skills (Goleman 1995). This diversification has been criticized for lacking empirical grounding (Waterhouse 2006), but can nonetheless be seen in discourse, even leading to framings like the “athletic intelligence” pursued by companies such as Boston Dynamics (Boston Dynamics 2025).

This emphasis on performance is also central to AI research. Planning and control theory emphasize achieving objectives efficiently with regard to a chosen metric (Kalman 1960; Kwakernaak and Sivan 1972; Lewis, Vrabie, and Syrmos 2012). Search and optimization generate behavior based on speed, cost, or accuracy (Fikes and Nilsson 1971; Nilsson 1980; Hsu, Campbell, and Hoane Jr 1995; Silver et al. 2016). In machine learning, ability is encoded in loss functions and performance metrics like accuracy, reward, or regret (Sra, Nowozin, and Wright 2011). These principles are also applied in models of language (Prince and Smolensky 1997) and sensorimotor behavior (Todorov 2004). Still, while performance-based views offer task-agnostic metrics, they often rely on narrow task formulations. An agent may excel at one task while failing at others. This raises questions about adaptability: should intelligence be measured by competence within tasks, or by breadth across them?

3.3 Generality – As General as Possible

A third perspective focuses on generality as a defining trait of intelligence, which refers to the capacity to perform across a wide range of tasks, environments, or goals. This view prioritizes flexibility and adaptability, sometimes bundled under related terms like robustness or generalization.

In AI, the desire for generality dates back to early calls for symbolic systems capable of cross-domain reasoning (McCarthy et al. 1955; Mitchell 1982). John McCarthy made it a central theme in his Turing Award Lecture (McCarthy 1987, presented in 1971). A similar idea also underlies modern machine learning, where generalization error measures model performance on unseen data (Nadeau and Bengio 1999). In robotics and control, desired generality can be operationalized as robustness to environmental changes (Dorato 1987; Yedavalli 2014; Mengers and Brock 2025; Battaje et al. 2026) or adaptability over time (Åström 1991; Landau et al. 2011). Across traditions, generality is treated as a desirable capability, while the rise of the buzzword “Artificial General Intelligence” (AGI) (Baum 2017) reflects growing interest across public and technical discourse.

Generality remains more emphasized in artificial than in biological intelligence research. In biology and psychology, simple generalization, such as stimulus generalization, is widespread across species (Guttman and Kalish 1956; Ghirlanda and Enquist 2003). However, more demanding notions of generality, such as flexibility across domains or novel goal adaptation, are also part of long-standing psychological and philosophical accounts of intelligence (Coelho Mollo 2024; Sternberg 2000). Thus, while generalization is not seen as exclusive to high intelligence, many theorists take some form of generality as a core diagnostic.

Importantly, generality does not require success in any environment or task. Rather, it involves competence across a broad enough set of contexts to support meaningful flexibility. Human intelligence, for example, is considered general not because it handles all possible tasks, but because it handles many different kinds (Lake et al. 2017). In sum, generality offers a way to evaluate intelligence not just by success on specific tasks, but by the scope and diversity of environments or goals an agent can handle. What constitutes sufficient scope remains contested across fields and applications.

4 AI Benchmarks and Conceptual Frameworks of Intelligence

This section examines how the conceptual frameworks outlined in Section 3 are operationalized in the contemporary landscape of AI evaluation. In this paper, benchmarks are not treated as the primary object of critique; rather, they serve as concrete sites where underlying conceptual commitments about intelligence are operationalized, making visible the implicit assumptions and exclusionary premises of different conceptual frameworks.

4.1 Anthropocentrism in AI Benchmarking

Although raw success metrics are central to most benchmarks, thus often invoking a conception of intelligence

grounded in ability, many current AI benchmarks also directly or indirectly encode human-likeness as an evaluative criterion. In benchmarks targeting computer use and interactive tasks, such as OSWorld (Xie et al. 2024) or Mind2Web (Deng et al. 2023), humans are either invoked explicitly as an anthropocentric ceiling on performance (Xie et al. 2024) or used as a reference for behavioral alignment, for instance by comparing an agent’s action traces to human demonstrations (Deng et al. 2023).

Beyond such explicit uses of humans as reference points, anthropocentric assumptions also shape benchmarks that rely on human preference judgments. When models are evaluated based on user preference, as in MT-Bench-101 (Bai et al. 2024) or the LMSYS Chatbot Arena (Elo) (Zheng et al. 2023), evaluation aggregates a heterogeneous set of behavioral properties. In addition to correctness, factors such as conversational ease, perceived helpfulness, expressivity, or confidence influence outcomes, implicitly judging how well a system aligns with human communicative norms. Recent work (Sypherd, Tang, and Belle 2025; Xiao et al. 2025) suggests that models can perform well under such evaluations by employing recognizable “social scripts,” including empathetic framing, first-person pronouns, or confident tone, which may increase perceived quality even when underlying reasoning is flawed.

Anthropocentrism is further reflected in the selection of benchmark tasks themselves. Common benchmarks focus on knowledge retrieval (Rein et al. 2024; Jin et al. 2019), dialogue (Bai et al. 2024; Zheng et al. 2023), image or video generation (Huang et al. 2024), software engineering (Jimenez et al. 2024), or interaction with graphical user interfaces (Deng et al. 2023; Xie et al. 2024). These tasks predominantly target forms of behavior that, outside of AI systems, are characteristic of humans, such as linguistic communication and interaction with digital tools and interfaces designed around human users and practices.

Even benchmarks for embodied agents (Li et al. 2024c,a), which could in principle emphasize domains not organized around human practices, such as ecological interaction, non-anthropomorphic locomotion, or environment-driven adaptation, are typically situated in human environments and framed around human activities, for example household manipulation or cleaning tasks.

4.2 Pure Ability in AI Benchmarking

Most AI benchmarks operationalize intelligence primarily through task success, accuracy, or completion rate, thereby conceptualizing intelligence as “pure ability”. Benchmarks such as GPQA (Rein et al. 2024) or PubMedQA (Jin et al. 2019) evaluate domain-specific knowledge, while SWE-bench (Jimenez et al. 2024) assesses competence in software engineering tasks. Benchmarking, by its nature, involves optimizing for performance on a defined task set; however, this focus on narrowly specified domains typically sidelines generality as an explicit evaluative goal. Notably, many such benchmarks are derived from tests originally designed to assess expert competence in humans. Examples include the bar exam or the American Invitational Mathematics Examination (AIME), which are used to certify specialized expertise

Conceptualization	Benchmark Instantiations	What Is It Excluding?	Normative Meaning	Epistemic Meaning
Anthropocentrism (Human-likeness)	Human preference benchmarks (MT-Bench-101, LMSYS Elo); imitation learning; Turing-style behavioral tests; tasks centered on human-computer interaction	Primarily non-human animals, plants, non-biological systems; non-human forms of cognition and embodiment	Speciesist hierarchy of worth; intelligence tied to human normativity; diminished moral and epistemic standing of non-humans	Essentialist and static view of intelligence centered on human cognition; limits cross-species comparison; obscures non-human forms of cognition and intelligence
Pure Ability (Optimal performance, rationality)	Task-specific benchmarks (GPQA, PubMedQA, SWE-bench); expert exams (bar exam, AIME); optimization metrics (accuracy, reward, efficiency)	Disabled individuals; those unable to perform tasks optimally; niche, non-normative skills and adaptive strategies	Ableist valuation of bodies and minds; intelligence associated with lack or deficit when optimality is unattainable	Conflation of intelligence with narrow task performance; neglect of adaptation under uncertainty; discourages study of intelligence in non-ideal conditions
Generality (Adaptability across environments and tasks)	Broad multi-task/-modal benchmarks (BIG-Bench, MMLU, MMMU); generalization and out-of-distribution tests (ARC-AGI)	Individuals whose cognition or agency depends on stable environments or support structures (e.g. those with severe dependencies); most current artificial systems	A weaker and more diffuse form of ableism; intelligence tied to adaptability rather than ideal performance	Shifts focus toward flexibility and ecological embeddedness; exposes gap between biological and artificial intelligence; highlights limits of current AI paradigms

Table 1: Conceptualizations of intelligence (Section 3), their instantiations in AI benchmarks (Section 4), and associated forms of exclusion, normativity, and epistemic meaning (Section 5).

in law or mathematics, respectively. When applied to AI systems (Bommarito II and Katz 2022; Li et al. 2024b), these assessments are often interpreted as indicators of broader “intelligence.” This shift mirrors historical tendencies in human psychometrics, where performance on tests of logical or analytical ability came to stand in for intelligence more generally. In this way, benchmarks grounded in expert performance instantiate an ability-based conception of intelligence that emphasizes correctness and competence within predefined tasks.

4.3 Generality in AI Benchmarking

Generality has become a central theme in contemporary AI discourse, most prominently through the notion of “artificial general intelligence” (AGI). Correspondingly, recent benchmarks increasingly aim to capture broader competence across tasks, domains, or modalities. Benchmarks such as MMLU (Hendrycks et al. 2021) or BIG-Bench (Srivastava et al. 2023) expand the range of evaluated subjects, while explicitly multimodal benchmarks like MMMU (Yue et al. 2024) broaden the types of input and reasoning required. These benchmarks often incorporate notions of out-of-distribution evaluation, seeking to assess performance on inputs not encountered during training. A complementary approach focuses more directly on adaptability to novel task settings. The Abstraction and Reasoning Corpus (ARC-AGI) (Chollet 2019), for example, is designed to minimize reliance on memorized knowledge and instead evaluates whether a system can infer and apply a simple rule in a previously unseen problem. Such benchmarks explicitly target generalization as a defining component of intelligence. Together, these approaches embed a generality-focused con-

ception of intelligence. At the same time, existing work has noted that practical limitations, such as data contamination or biases in task and dataset construction, can constrain the extent to which such benchmarks capture generality in practice (Cheng et al. 2025). Nonetheless, their design reflects an increasing emphasis on adaptability and breadth as central markers of intelligence in current AI evaluation.

5 Haves and Have Nots: Who or What Gets Excluded?

The previous section has shown that conceptualizations of intelligence matter for practical reasons, as they manifest in benchmarks used in AI. Furthermore, this paper also discussed in Section 2 how technoscience is not free of bias and exclusion. The very concept of intelligence relies on epistemic exclusion, with haves and have nots. This section explores the exclusionary practices within the framework of the three conceptualizations introduced in Section 3, by first defining the descriptive exclusion (what is counted as intelligence within a framework), second normative meaning (forms of oppression and injustice associated with the exclusion), and epistemic meaning (how this exclusion is related to knowledge production in the science of intelligence). An overview of our discussion is provided in Table 1 and is further explained in the next paragraphs.

5.1 Anthropocentrism and Speciesism

What is it excluding? The anthropocentric perspective on intelligence could be seen as a form of speciesism. Broadly speaking, speciesism refers to favoring, or discriminating one species over another (Horta 2010), in which one is seen as “lesser” than the other, in this case humans trumping over

non-humans, such as plants and non-human animals. This reflects an essentialist view of intelligence, treating it as an inherent trait of a species (e.g., humans) rather than as a set of behaviors that are performed. According to this perspective, individuals are born either in or out of intelligence due to their species.² The speciesist reading, however, does not exhaust what anthropocentrism excludes. As Rhee (2018) argues, the figure of “the human” that AI and robotics reproduce is itself a normative construction rather than a neutral reference point, historically drawn narrowly enough to exclude those whose labor, embodiment, or ways of communicating fall outside it. Anthropocentrism therefore operates along two axes: it excludes non-humans by definition, and it disadvantages those humans who do not match the particular conception of “humanness” being invoked.

Normative Meaning The epistemic exclusion of other species in this definition of intelligence also has societal implications, particularly regarding animal rights. Several authors have argued that speciesism is related to hierarchies of perceived intelligence and cognition, which, in turn, could affect how animals are being treated. Singer (2009) for instance contends that speciesism and mistreating animals are related to the understanding that human beings have better cognitive capacities.³ Other literature makes a similar claim, that speciesism relies on othering animals and reinforces a human-animal dualistic divide on the basis of cognitive abilities (Salter 2015). Moreover, Coeckelbergh and Gunkel (2014) also refer to hierarchy, though they do not articulate a dualistic divide, but rather a ranking among animals, where those who are supposedly closer to human intelligence being perceived as “higher” in a normative “ranking” than others. Nonetheless, Coeckelbergh and Gunkel (2014) mention that while we attribute more positive attributes to animals who are seemingly more intelligent and comparable to humans than other animals, it does not always translate into more ethical behavior towards these animals. For instance, it seems rather arbitrary according to the authors that pigs are being eaten while other animals like cats or dogs are used for companion purposes.

Epistemic Meaning To date, there is little principled justification why other types of species, for instance plant intelligence, should be excluded from the definition of intelligence, as “[...] plants must acquire the resources they need to grow, they need to deal with predators and disease and find mates. Instead of movement, the competitive fight for the essentials, light, minerals and water, led instead to fights over space” (Trewavas 2017, p. 1). Trewavas’s explanation of plant intelligence therefore points in the direction of Darwinian fitness (Cvrčková, Lipavská, and Žárský 2009), and

²In this context, being born into intelligence does not mean that human intelligence cannot be influenced by environmental factors. Rather, it means that non-human animals are by definition excluded from intelligence with this perspective.

³Cognition and intelligence are not the only reference points for how humans treat non-human animals. Among others, the criteria if they feel pain or are able to suffer are other questions taken into account when discussing animal ethics.

moves away from so-called “brain chauvinism” in which intelligence does not need a brain or nervous system to be classified as such (Trewavas 2017).

Similarly, Coelho Mollo (2024) argues for a more inclusive framework of intelligence, as it would benefit the science of intelligence. According to him, this speciesism, and specifically anthropocentrism, is problematic for several reasons. First, by definition, anthropocentrism limits “[...] without justification the realm of what is intelligent to the specific cognitive and motor abilities, needs, interests, and ecological embeddedness of a single biological species” (Coelho Mollo 2024, p. 707). According to him, not only does it constrain without good reason the possibilities for studying different forms of cognition and motor skills found in non-humans, furthermore, anthropocentrism hinders the possibility of studying general principles of intelligence, found across species, or as he mentions: “anthropocentrism denies that there can be a cross-disciplinary notion of intelligence covering also the behaviors and abilities of non-human systems—biological and artificial—that ethologists, animal cognition researchers, comparative psychologists, and AI researchers are interested in” (Coelho Mollo 2024, p. 707). This criticism against speciesism, and specifically anthropocentrism, is therefore rooted in the study of intelligence. By defining intelligence as something humans possess, the author argues it will not benefit, but rather limit, the field of intelligence research. Coelho Mollo (2024) in other words refers to unfounded epistemic exclusion: acknowledging that other species might have different kinds of intelligence that ultimately could be beneficial to study, even though they are excluded in some research strands of intelligence research. This type of epistemic exclusion therefore employs a very generalist function: not behavior or interspecies variability, rather the species itself as a whole is seen as incapable of being intelligent. This definition of intelligence could therefore be seen as a static and essentialist view of intelligence: you are born into a species, and therefore born into intelligence.

5.2 Ability and Ableism

What is it excluding? Unlike speciesism, defining intelligence in terms of pure ability does not refer to humans per se, as it is about optimal behavior, i.e., the most efficient way of executing a certain task. The question then becomes who or what is then able to most optimally execute a certain task? The answer to this question of course depends on the task at hand, but it does presuppose being *able* to perform this task. Any type of bodily or cognitive impairment that hinders optimal behavior therefore falls out of this definition of intelligence. For instance, singing a melody optimally, requires one to not be deaf. In other words, intelligence as optimality represents a differentiation of bodies, in which performing “optimally” ultimately categorizes able-bodied haves and non-abled have nots. This is also apparent in the historical context of this research strand in the science of intelligence, as intelligence testing has a history of quantifying normality, excellence, and disorders. In other words, conceptualizing intelligence as optimality manifests in testing and differentiating individuals based on their ability.

Normative Meaning An ableist approach to intelligence puts emphasis on being capable to perform certain tasks optimally. Those who cannot perform optimal behavior would display what Scully calls “inbuilt epistemic limitation” (Scully 2018, p. 110), as their bodies are defined by lack. This could be seen as a purely descriptive aspect of intelligence: e.g., some people are able to see without aids, some are visually impaired, some people are blind, which makes certain tasks involving vision difficult for those who are lacking it. However, the implications of this classification rarely remain categorical and descriptive. Being able to perform a task is seen as beneficial. The problem with this perspective of intelligence is not that it seeks to excel at certain tasks in an effective manner, but rather that it could exclude certain niche skills and knowledge that non-abled individuals have to develop in order to solve tasks. Neglecting this type of knowledge also has consequences for the study of intelligence as is explained next.

Epistemic Meaning While an ableist approach of intelligence excludes disabled people from the definition, and therefore has ethical implications, it should be questioned what it means for the science of intelligence. After all, at face value the “most efficient” way to perform a task could be seen as beneficial. From a science of intelligence perspective, the pursuit of optimal performance is not inherently problematic. In fact, we can often measure the performance of a given organism against a chosen metric or instruct machines to optimize it. However, this optimization-centric view limits the science of intelligence in a number of ways. First, there are feasibility issues with the assumption that intelligence consists of optimal responses to well-defined problems presumes a static and fully knowable environment. In reality, most environments are dynamic, noisy, and uncertain. Perceiving and acting under such conditions requires not perfect information, but strategies for coping with ambiguity, adaptability over time, and robust generalization. Second, there are implications for the types of machines built because this paradigm tends to produce systems with highly specialized capabilities what Burema (2021) refers to—albeit with ethically questionable terminology—as “idiot savant” (Burema 2021, p. 154) machines: entities that excel at narrow tasks but lack the general intelligence required to navigate diverse and unpredictable scenarios. If our scientific understanding of intelligence is shaped by these narrow benchmarks, we risk conflating intelligence with performance under idealized conditions.

As Moravec’s Paradox (Moravec 1988) highlights, there is a misalignment between what is cognitively difficult for humans and what is difficult for machines. We tend to emphasize tasks that are cognitively hard for us (e.g., abstract reasoning, logic) over tasks that are actually difficult for machines, such as sensorimotor adaptability in noisy, dynamic environments. This discrepancy reveals the limitations of optimization-based approaches, which may obscure rather than illuminate the true nature of intelligence and constrain the kinds of questions researchers ask about it.

Relatedly, expanding our understanding of intelligence to individuals who are not performing behavior optimally

would benefit the science of intelligence, by providing insights into how intelligence operates in less-than-ideal conditions. For instance, Scully (2018) mentions wheelchair users have unique lived experiences and, therefore, possess specific kinds of knowledge that non-wheelchair users do not have. A wheelchair user might not display optimal behavior in terms of getting from point A to point B due to factors like poor urban planning or construction work. Yet, they can be able to succeed in navigating everyday life and overcoming obstacles. From a science of intelligence perspective, this illustrates how even when faced with additional constraints, individuals can still engage in “successful” behavior. This insight underscores the need to broaden the scientific view of intelligence, taking into account the adaptability required to navigate the messiness of everyday life.

5.3 Generality and Ableism

What is it excluding? Generality could be regarded as the most inclusive form of intelligence. It does not hold an essentialistic perspective of intelligence that for instance anthropocentrism does embed. In other words, it is not something an individual “has” like a fixed entity that one is born into, but it refers to behavior. It includes various species, as well as individuals that might perform more or also *less* “optimally” on a given task. The main criterion for intelligence in this case is being able to adapt to ever-changing circumstances, which is something general populations of varying biological organisms are able to perform fairly well (Roli, Jaeger, and Kauffman 2022). Though seemingly inclusive, this statement is not free of exclusion. It excludes those who are not able to adapt to changing environments, who are typically severely disabled individuals.

Normative Meaning This therefore embeds certain notions of ableism, as some being able to be better at certain behaviors than others. Nonetheless, the difference between this form of ableism in generality and in pure ability has to do with outliers: being optimal at a given task gives space for an “ideal.” This by definition discards every individual who is not able to perform optimally. By contrast, most individuals across species are able to perform some level of general intelligence. While there exist differences between different humans as well as animals, generality seems to be present in all of them to a large degree; especially when we account for different ecological niches. It primarily excludes those who have extreme dependencies, and therefore might be seen as less exclusionary than the other conceptual frameworks discussed in this paper.

Epistemic Meaning As noted earlier, generality appears to be a widespread feature of biological life, particularly when considered in relation to an organism’s ecological niche. Within their specific environmental contexts, most organisms demonstrate an ability to adapt to changing conditions (Sexton et al. 2017). In contrast, artificial systems rarely exhibit such generality. This is especially true for embodied agents operating in the real world, but even digital systems with less broad application domains, such as vacuum robots or language models like ChatGPT, struggle to generalize beyond their training context.

	<i>On the Conception of Intelligence</i>
Q1	What behaviors or capacities does this benchmark measure?
Q2	Which prior benchmarks or research traditions does this benchmark build on?
Q3	What assumptions about intelligence does this benchmark inherit from those traditions?
Q4	Who or what serves as the reference point for intelligent behavior in this benchmark, e.g., a human expert, a specific population, or an abstract performance standard?
	<i>On Exclusion</i>
Q5	Which agents, species, or groups are excluded from evaluation entirely by this benchmark?
Q6	Which environments/tasks are excluded from evaluation entirely by this benchmark?
Q7	Among the agents this benchmark is designed to evaluate, which groups or individuals are likely to perform worse due to the specific tasks chosen, for instance due to physical ability, language, or socio-demographic characteristics?
	<i>On Scope</i>
Q8	What aspects of intelligent behavior does this benchmark not measure?

Table 2: A structured set of questions to foster reflexivity in the design and selection of AI benchmarks.

This inclusion of almost all natural intelligence while excluding synthetic systems suggests a gap in our understanding of how generality arises. As Feynman (1988) famously implied, if we truly understood a phenomenon, we should be able to recreate it. This gap resonates with the broader insight of Moravec’s paradox (Moravec 1988): that intelligence as exhibited in the real world may rely on forms of adaptation and embodiment that have been undervalued or overlooked in traditional research and societal interpretations of “intelligence.”

6 Discussion: Fostering Reflexivity in Benchmark Design

The previous sections illustrate how intelligence research embeds ableist and speciesist notions of intelligence, and how existing benchmarks fall into particular conceptual traditions without always making this explicit. In order to facilitate critical thinking in the design and selection of AI benchmarks, we propose a set of questions in Table 2 that help uncover which normativities are embedded within a given benchmark’s design. The goal is not inclusivity for its own sake, but rather to make taken-for-granted assumptions about intelligence explicit, enabling researchers to make choices critically and deliberately.

The three conceptual frameworks discussed throughout this paper provide the analytical lens through which these questions are organized. As Table 1 shows, each framework carries a characteristic pattern of exclusion: anthropocentrism excludes along species lines, giving rise to speciesist assumptions, pure ability excludes along lines of optimal performance, giving rise to ableist assumptions, and generality, while the most inclusive of the three, still excludes those whose cognition depends on stable environments or support structures. The questions in Table 2 are designed to surface these patterns in any given benchmark, whether or not the designers were aware of them.

The first set of questions addresses the conception of intelligence embedded in the benchmark: what is being measured (Q1), which prior traditions it builds on and what as-

sumptions it inherits from them (Q2 and Q3), and who or what serves as the reference point for intelligent behavior (Q4). The second set of questions targets exclusion specifically, asking which agents, species, or environments fall outside the benchmark’s scope (Q5 and Q6), and which groups within the target population are systematically disadvantaged by the task design (Q7). The final question concerns scope more broadly: what aspects of intelligent behavior the benchmark does not measure (Q8).

Similar to reflexivity in ethnographic fieldwork, this set of questions is meant as an exercise to position oneself in relation to the object of study, questioning methodological assumptions, category constructions, and power relations. While reflexivity in ethnographic fieldwork focuses on the researcher’s identity and positionality, at its core the concept illustrates how method and outcome are shaped by their broader cultural environment. We see reflexivity in benchmark design as a practice of making explicit, and documenting, the assumptions about intelligence and exclusion that are inscribed into the artifact. To illustrate what these questions yield in practice, we work through them for MMLU (Hendrycks et al. 2021) in Table 3. It shows how the scope of this benchmark is limited to groups and species who communicate with written language, and has a bias towards US-based knowledge. Moreover, it does not include, among others, sensorimotor competence or non-human agents as reference points for intelligence.

7 Conclusion

This paper focused on the meaning of different conceptual frameworks of intelligence, and how that translates to benchmark design. As Table 1 shows, each conceptualization carries distinct forms of exclusion and epistemic consequences, highlighting why reflexivity in AI and intelligence research is crucial. While past literature has often focused on performance metrics, IQ scores and intelligence testing (Fischbach, Vandemeulebroucke, and Van Wynsberghe 2023; Alfano, Holden, and Conway 2017), or the misuse of the term by actors in AI (Cave 2020; Gebru and Torres 2024),

<i>On the Conception of Intelligence</i>	
Q1	MMLU measures world knowledge and problem-solving ability across 57 subjects, presented as four-option multiple-choice items.
Q2	Its items are drawn from human standardized testing—practice questions for the GRE and the USMLE, undergraduate course material, and items modeled on Advanced Placement examinations—while the benchmark positions itself against earlier NLP benchmarks whose rapid saturation it takes as evidence that they failed to capture language understanding.
Q3	From the testing tradition it inherits the premise that aggregating performance across many knowledge domains indexes a general underlying capacity.
Q4	Its reference point is explicitly human and stratified: unspecialized annotators reach 34.5% accuracy, while “expert-level” performance is set at approximately 89.8%, derived from 95th-percentile human test-taker scores and, where those were unavailable, from the authors’ estimate. A benchmark framed in terms of breadth is thus anchored to an anthropocentric ceiling and an ability-based account of what that breadth is for.
<i>On Exclusion</i>	
Q5	Only agents that process written language can be evaluated, and therefore groups and species that do not communicate with written language are excluded.
Q6	The format excludes embodied, multimodal, and open-ended settings.
Q7	The benchmark is English-only, and a substantial subset of its subjects is keyed to US institutions—US History, US Foreign Policy, Government and Politics, Professional Law, Professional Accounting, and Professional Medicine each presuppose familiarity with US curricula, jurisprudence, tax code, or licensing examinations. Performance on these subjects tracks familiarity with one national context rather than the general capacity the benchmark reports. Groups who would perform worse therefore include subjects who are not native English speakers and from the US.
<i>On Scope</i>	
Q8	MMLU does not measure adaptation under uncertainty, sensorimotor competence, sustained interaction, or knowledge that resists reduction to a single correct option.

Table 3: A worked example of the reflexivity questions in Table 2 for the MMLU benchmark (Hendrycks et al. 2021).

we have taken a different approach: first, discussing epistemic exclusion and its meaning across three conceptual frameworks of intelligence—anthropocentrism, pure ability, and generality—and second, proposing a set of questions to foster epistemic reflexivity when designing AI benchmarks.

Regarding the former, we argue that these conceptual frameworks often privilege able-bodied humans as the “haves” of intelligence, while systematically excluding others, such as disabled individuals or non-human species. Intelligence is not only a descriptive concept, but also a normative one. It is not a neutral or value-free term, but one that carries significant weight in both scientific and societal contexts. As previous scholars have shown, this concept has historically been tied to eugenics (Gebru and Torres 2024).

Regarding the second aim of the paper, we build on the idea that conceptualizations of intelligence are related to technological development. STS research has shown that assumptions about the world matter and carry a morally performative element when designing technologies (e.g., Verbeek 2006), which also shows in the benchmarks used in AI development. Therefore, questioning these assumptions of intelligence is not just a philosophical exercise, but gives insights on practical engineering and decision-making processes. To elaborate, prioritizing human-likeness may unnecessarily constrain the design space, even when non-human-like systems might be more effective or desirable in certain contexts. Similarly, optimization-centered approaches tend to favor narrow, specialized capabilities, limiting the development of systems that are robust, general, and adaptable. Accommodating alternative conceptions of intel-

ligence in AI benchmarks is feasible, as seen in the explicit inclusion of animal behavior (Crosby et al. 2020). We therefore argue for epistemic reflexivity, in which design choices are made explicit. We proposed a set of questions for AI developers to document their thought process on how definitions of intelligence are inscribed into benchmarks.

Though the main objective of this paper does not argue how the science of intelligence should “progress,” it should be noted that these exclusionary meanings may hinder the science of intelligence itself. As we have argued, excluding certain perspectives—whether from other species or disabled individuals—may cause us to overlook relevant phenomena. Interestingly, the least exclusionary conceptualization, i.e., generality, is increasingly gaining traction in AI research. While this shift is not necessarily driven by an inclusivity agenda, it may incidentally help broaden the scope of inquiry. By allowing a wider range of perspectives and abilities to be included, generality may help us uncover principles of intelligence across contexts, species, and forms.

Ultimately, this paper does not argue that methodologies based on any particular conceptualization of intelligence must be abandoned. Intelligence as a concept by definition would always have an exclusionary element, as it inherently distinguishes between “haves” and “have nots,” where even generality excludes certain individuals. This is not a flaw; this is a feature. Instead of advocating for an all-encompassing and completely inclusive definition of intelligence, we argue for a conception of intelligence that takes its exclusions seriously. We must therefore continue to ask the question: Whose “intelligence” are we building?

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/2 “Science of Intelligence” – project number 390523135. We thank our colleagues at the Science of Intelligence cluster for valuable discussions and feedback on this work.

References

- Akrich, M. 1992. The de-scription of technical objects. In Bijker, W.; and Law, J., eds., *Shaping technology/building society. Studies in sociotechnical change*, 205–224. MIT Press.
- Alfano, M.; Holden, L.; and Conway, A. 2017. Intelligence, Race, and Psychological Testing. In Zack, N., ed., *The Oxford Handbook of Philosophy and Race*, 474–486. Oxford University Press.
- Åström, K. J. 1991. Adaptive control. In *Mathematical System Theory: The Influence of RE Kalman*, 437–450. Springer.
- Bai, G.; Liu, J.; Bu, X.; He, Y.; Liu, J.; Zhou, Z.; Lin, Z.; Su, W.; Ge, T.; Zheng, B.; et al. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, volume 1, 7421–7454.
- Battaje, A.; Bernhard, M.; Mengers, V.; and Brock, O. 2026. Building Generalization Into Behavior Generation Via Adaptive Compositions of Regularities. *arXiv preprint arXiv:2605.31110*.
- Baum, S. 2017. A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Working Paper*, 17–1.
- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. ACM. ISBN 978-1-4503-8309-7.
- Berkeley, I. S. 2019. The curious case of connectionism. *Open Philosophy*, 2(1): 190–205.
- Bijker, W. E. 2010. How is technology made?—That is the question! *Cambridge Journal of Economics*, 34(1): 63–76.
- Billard, A. G.; Calinon, S.; and Dillmann, R. 2016. Learning from humans. *Springer Handbook of Robotics*, 1995–2014.
- Blili-Hamelin, B.; Hancox-Li, L.; and Smart, A. 2024. Unsocial Intelligence: An Investigation of the Assumptions of AGI Discourse. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 141–155.
- Bommarito II, M.; and Katz, D. M. 2022. GPT takes the bar exam. *arXiv preprint arXiv:2212.14402*.
- Bordes, F.; Ross, C.; Kao, J. T.; Spiliopoulou, E.; and Williams, A. 2025. Eval factsheets: A structured framework for documenting AI evaluations. *arXiv preprint arXiv:2512.04062*.
- Boston Dynamics. 2025. Atlas® and beyond: the world's most dynamic robots. <https://bostondynamics.com/atlas/>. Accessed: 2025-04-02.
- Bouchard Jr, T. J. 1998. Genetic and environmental influences on adult intelligence and special mental abilities. *Human Biology*, 257–279.
- Brock, O. 2024. Intelligence as Computation. In *IOP Conference Series: Materials Science and Engineering*, volume 1321:1, 012001. IOP Publishing.
- Burema, D. 2021. *Engineering elder care: An analysis of conceptual premises and biases of social robots in elder care*. Ph.D. thesis, Universität Bremen.
- Burema, D. 2025. The challenges of being an in-house AI ethicist and how to overcome them. *Journal of Responsible Innovation*, 12(1): 2445322.
- Cave, S. 2020. 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*, 29–35. ACM. ISBN 978-1-4503-7110-0.
- Cheng, Z.; Wahnig, S.; Gupta, R.; Alam, S.; Abdullahi, T.; Ribeiro, J. A.; Nielsen-Garcia, C.; Mir, S.; Li, S.; Orender, J.; et al. 2025. Benchmarking is Broken—Don't Let AI be its Own Judge. *Advances in Neural Information Processing Systems*.
- Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2025. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11): 1684–1704.
- Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.
- Coeckelbergh, M.; and Gunkel, D. J. 2014. Facing animals: A relational, other-oriented approach to moral standing. *Journal of Agricultural and Environmental Ethics*, 27: 715–733.
- Coelho Mollo, D. 2024. Intelligent Behaviour. *Erkenntnis*, 89(2): 705–721.
- Crosby, M.; Beyret, B.; Shanahan, M.; Hernández-Orallo, J.; Cheke, L.; and Halina, M. 2020. The animal-AI testbed and competition. In *Conference on Neural Information Processing Systems: Competition and Demonstration Track*, 164–176.
- Cvrčková, F.; Lipavská, H.; and Žárský, V. 2009. Plant intelligence: why, why not or where? *Plant Signaling & Behavior*, 4(5): 394–399.
- De Waal, F. 2016. *Are we smart enough to know how smart animals are?* WW Norton & Company.
- Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36: 28091–28114.
- Dorato, P. 1987. A historical review of robust control. *IEEE Control Systems Magazine*, 7(2): 44–47.
- Feynman, R. 1988. “What I cannot create, I do not understand”. Written on his blackboard at Caltech at the time of his death.

- Fikes, R. E.; and Nilsson, N. J. 1971. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2(3-4): 189–208.
- Fischbach, F.; Vandemeulebroucke, T.; and Van Wynsberghe, A. 2023. Mind who’s testing: Turing tests and the post-colonial imposition of their implicit conceptions of intelligence. *AI & Society*.
- Gardner, H.; and Hatch, T. 1989. Multiple Intelligences Go to School: Educational implications of the theory of multiple intelligences. *Educational Researcher*, 18(8): 4–10.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Gebru, T.; and Torres, É. P. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*.
- Ghirlanda, S.; and Enquist, M. 2003. A century of generalization. *Animal Behaviour*, 66(1): 15–36.
- Goleman, D. 1995. *Emotional Intelligence: Why It Can Matter More Than IQ*. Bantam Books.
- Gonzalez, G. 1979. The Historical Development of the Concept of Intelligence. *Review of Radical Political Economics*, 11(2): 44–54.
- Guttman, N.; and Kalish, H. I. 1956. Discriminability and stimulus generalization. *Journal of Experimental Psychology*, 51(1): 79.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. In *The Ninth International Conference on Learning Representations*.
- Horta, O. 2010. What is speciesism? *Journal of Agricultural and Environmental Ethics*, 23: 243–266.
- Hsu, F.-h.; Campbell, M. S.; and Hoane Jr, A. J. 1995. Deep Blue system overview. In *Proceedings of the 9th International Conference on Supercomputing*, 240–244.
- Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; and Chanpaisit, N. e. a. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.
- Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. R. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations*.
- Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; and Lu, X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2567–2577.
- Kalman, R. E. 1960. Contributions to the Theory of Optimal Control. *Boletín de la Sociedad Matemática Mexicana*, 5(2): 102–119.
- Kwakernaak, H.; and Sivan, R. 1972. *Linear optimal control systems*, volume 1. Wiley-Interscience.
- Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40: e253.
- Landau, I. D.; Lozano, R.; M’Saad, M.; and Karimi, A. 2011. *Adaptive control: algorithms, analysis and applications*. Springer Science & Business Media.
- Legg, S.; and Hutter, M. 2007. A Collection of Definitions of Intelligence. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, 17–24.
- Lewis, F. L.; Vrabie, D.; and Syrmos, V. L. 2012. *Optimal control*. John Wiley & Sons.
- Li, C.; Zhang, R.; Wong, J.; Gokmen, C.; Srivastava, S.; Martín-Martín, R.; Wang, C.; Levine, G.; Ai, W.; Martinez, B.; et al. 2024a. BEHAVIOR-1K: A Human-Centered, Embodied AI Benchmark with 1, 000 Everyday Activities and Realistic Simulation. *CoRR*.
- Li, J.; Beeching, E.; Tunstall, L.; Lipkin, B.; Soletskyi, R.; Huang, S.; Rasul, K.; Yu, L.; Jiang, A. Q.; Shen, Z.; et al. 2024b. NuminaMath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face Repository*, 13(9): 9.
- Li, M.; Zhao, S.; Wang, Q.; Wang, K.; Zhou, Y.; Srivastava, S.; Gokmen, C.; Lee, T.; Li, E. L.; Zhang, R.; et al. 2024c. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37: 100428–100534.
- Lindlof, T. R.; and Taylor, B. C. 2017. *Qualitative communication research methods*. Sage publications.
- Mackintosh, N. J. 2011. History of theories and measurement of intelligence. *The Cambridge Handbook of Intelligence*, 3–19.
- McCarthy, J. 1987. Generality in Artificial Intelligence [Postscript]. *Communications of the ACM*, 30(12): 1030–1035. Originally presented in 1971.
- McCarthy, J.; Minsky, M. L.; Rochester, N.; and Shannon, C. E. 1955. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Technical report, Dartmouth College.
- Mei, Q.; Xie, Y.; Yuan, W.; and Jackson, M. O. 2024. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9): e2313925121.
- Mengers, V.; and Brock, O. 2025. No Plan but Everything Under Control: Robustly Solving Sequential Tasks with Dynamically Composed Gradient Descent. In *IEEE International Conference on Robotics and Automation*, 90–96.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.
- Mitchell, T. M. 1982. Generalization as search. *Artificial Intelligence*, 18(2): 203–226.
- Moravec, H. 1988. *Mind children: The future of robot and human intelligence*. Harvard University Press.

- Nadeau, C.; and Bengio, Y. 1999. Inference for the generalization error. *Advances in Neural Information Processing Systems*, 12.
- Nilsson, N. J. 1980. *Principles of artificial intelligence*. Tioga Press.
- Peterson, J. 1925. *Early conceptions and tests of intelligence*. World Book Company.
- Pfeifer, R.; and Gómez, G. 2009. Morphological computation—connecting brain, body, and environment. *Creating Brain-Like Intelligence: From Basic Principles to Complex Intelligent Systems*, 66–83.
- Pinch, T. J.; and Bijker, W. E. 1984. The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, 14(3): 399–441.
- Polderman, T. J.; Benyamin, B.; De Leeuw, C. A.; Sullivan, P. F.; Van Bochoven, A.; Visscher, P. M.; and Posthuma, D. 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7): 702–709.
- Prince, A.; and Smolensky, P. 1997. Optimality: From Neural Networks to Universal Grammar. *Science*, 275(5306): 1604–1610.
- Rajaraman, N.; Yang, L.; Jiao, J.; and Ramchandran, K. 2020. Toward the fundamental limits of imitation learning. *Advances in Neural Information Processing Systems*, 33: 2914–2924.
- Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- Rhee, J. 2018. *The robotic imaginary: The human and the price of dehumanized labor*. University of Minnesota Press.
- Roli, A.; Jaeger, J.; and Kauffman, S. A. 2022. How organisms come to know the world: Fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, 9: 806283.
- Rosenblatt, F. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(6): 386.
- Russell, S. J. 1997. Rationality and Intelligence. *Artificial Intelligence*, 94(1-2): 57–77.
- Russell, S. J.; Norvig, P.; and Davis, E. 2010. *Artificial intelligence: a modern approach*. Prentice Hall series in artificial intelligence. Prentice Hall, 3rd edition. ISBN 978-0-13-604259-4.
- Salter, C. 2015. Animals and war: anthropocentrism and technoscience. *NanoEthics*, 9: 11–21.
- Scully, J. L. 2018. From “she would say that, wouldn’t she?” to “does she take sugar?” epistemic injustice and disability. *International Journal of Feminist Approaches to Bioethics*, 11(1): 106–124.
- Sexton, J. P.; Montiel, J.; Shay, J. E.; Stephens, M. R.; and Slatyer, R. A. 2017. Evolution of ecological niche breadth. *Annual Review of Ecology, Evolution, and Systematics*, 48(1): 183–206.
- Shettleworth, S. J. 2009. *Cognition, evolution, and behavior*. Oxford University Press.
- Shevlin, H.; Vold, K.; Crosby, M.; and Halina, M. 2019. The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge. *EMBO Reports*, 20(10): e49177.
- Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587): 484–489.
- Simon, H. A.; and Newell, A. 1971. Human Problem Solving: The State of the Theory in 1970. *American Psychologist*, 26(2): 145.
- Singer, P. 2009. Speciesism and Moral Status. *Metaphilosophy*, 40(3-4): 567–581.
- Sokol, A.; Daly, E.; Hind, M.; Piorkowski, D.; Zhang, X.; Moniz, N.; and Chawla, N. 2025. BenchmarkCards: Standardized Documentation for Large Language Model Benchmarks. *Advances in Neural Information Processing Systems*, 38.
- Sra, S.; Nowozin, S.; and Wright, S. J. 2011. *Optimization for machine learning*. MIT Press.
- Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.
- Sternberg, R. J. 2000. *Handbook of intelligence*. Cambridge University Press.
- Sypherd, C.; Tang, W.; and Belle, V. 2025. Breaking the illusion: Revisiting LLM anthropomorphism. In *The 4th International Conference on Human and Artificial Rationalities*.
- Todorov, E. 2004. Optimality principles in sensorimotor control. *Nature Neuroscience*, 7(9): 907–915.
- Trewavas, A. 2017. The foundations of plant intelligence. *Interface Focus*, 7(3): 1–18.
- Turing, A. 1950. Computing Machinery and Intelligence. *Mind*, 433–460.
- Varela, F. J.; Thompson, E.; and Rosch, E. 1992. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.
- Verbeek, P.-P. 2006. Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, 31(3): 361–380.
- Waterhouse, L. 2006. Multiple intelligences, the Mozart effect, and emotional intelligence: A critical review. *Educational Psychologist*, 41(4): 207–225.
- Xiao, Y.; Ng, L. H. X.; Liu, J.; and Diab, M. 2025. Humanizing machines: Rethinking llm anthropomorphism through a multi-level framework of design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3331–3350.
- Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T. J.; Cheng, Z.; Shin, D.; Lei, F.; et al. 2024. Oworld:

Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37: 52040–52094.

Yedavalli, R. K. 2014. *Robust control of uncertain dynamic systems*, volume 10. Springer.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zare, M.; Kebria, P. M.; Khosravi, A.; and Nahavandi, S. 2024. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Conference on Neural Information Processing Systems: Datasets and Benchmarks Track*.