Realist and Pluralist Conceptions of Intelligence and Their Implications on AI Research

Ninell Oldenburg [1], **Ruchira Dhar** [2], **Anders Søgaard**[1,2]

[1]Department of Philosophy, University of Copenhagen
[2]Department of Computer Science, University of Copenhagen
niol@hum.ku.dk, rudh@di.ku.dk, soegaard@di.ku.dk

Abstract

In this paper, we argue that current AI research operates on a spectrum between two different underlying conceptions of intelligence: *Intelligence Realism*, which holds that intelligence represents a single, universal capacity measurable across all systems, and *Intelligence Pluralism*, which views intelligence as diverse, context-dependent capacities that cannot be reduced to a single universal measure. Through an analysis of current debates in AI research, we demonstrate how the conceptions remain largely implicit yet fundamentally shape how empirical evidence gets interpreted across a wide range of areas. These underlying views generate fundamentally different research approaches across three areas. Methodologically, they produce different approaches to model selection, benchmark design, and experimental validation. Interpretively, they lead to contradictory readings of the same empirical phenomena, from capability emergence to system limitations. Regarding AI risk, they generate categorically different assessments: realists view superintelligence as the primary risk and search for unified alignment solutions, while pluralists see diverse threats across different domains requiring context-specific solutions. We argue that making explicit these underlying assumptions can contribute to a clearer understanding of disagreements in AI research.

Introduction

There is a growing debate among researchers on the nature and potential risks of AI, as well as the importance of respective research programs. On one hand, researchers warn that building bigger models may eventually lead to human-like, superhuman, or "transformative" and possibly dangerous levels of intelligence (Amodei et al. 2016; Bengio et al. 2025; Phan et al. 2025; Hendrycks, Mazeika, and Woodside 2023). On the other hand, scholars argue that simply creating larger models is unlikely to yield human-like cognition (Raji et al. 2021; Mitchell 2024; Johnson et al. 2024).

What is intriguing about this debate is that researchers are observing the same empirical data, model architectures, scaling laws, and model performance, while holding on to profoundly different interpretations, e.g., scaling laws as a sign of emerging superintelligence vs. task-specific improvements. In this article, we suggest that this dichotomy can be traced back to diverging beliefs on what intelligence is. On one hand, intelligence realists hold that intelligence represents a single, universal capacity measurable across all systems. On the other hand, intelligence pluralists see intelligence as diverse, context-dependent capacities that cannot be reduced to a single universal measure. This has first-order implications for how we understand "artificial" intelligence, what we consider useful and dangerous, how we test, evaluate, and interpret systems, and how we build them. Different interpretations of empirical scaling laws, i.e., whether they represent convergence toward universal intelligence or domain-specific optimization, depend on which philosophical framework researchers inherently hold.

Our framework contributes to the philosophy of science by demonstrating how paradigmatic commitments shape observation and interpretation in contemporary AI research. Following Kuhn (1962)'s theory that theoretical frameworks determine what counts as significant evidence, we show how realist and pluralist assumptions function as competing paradigms that make identical empirical results (scaling curves, benchmark performance, system failures) intelligible in fundamentally different ways. This extends Hanson (1958)'s theory-ladenness of observation to AI: what researchers "see" in GPT-4's performance depends on their prior commitments about intelligence's nature. By making these implicit paradigms explicit, we enable more productive scientific discourse and clearer identification of genuine empirical disagreements vs. philosophical differences.

The remainder of this paper is structured as follows. We first present the features of each view with examples from the literature. We acknowledge that this conceptualization should be seen as a continuum rather than a hard classification before describing the implications these positions have on AI research methodology and AI alignment.

Our contribution is thereby not to invent a new distinction but to provide a *targeted synthesis* that contextualizes longstanding debates in psychology, philosophy, and cognitive science within contemporary AI research. What is novel is demonstrating how these debates *structure AI research in ways researchers often do not recognize*. By making explicit the implicit philosophical commitments underlying methodological choices, interpretive disagreements, and alignment strategies, we provide a vocabulary for clearer scientific discourse and more productive policy debates.

Dimension	Realist Indicators	Intermediate/Mixed	Pluralist Indicators
Ontology	Intelligence as singular capacity; seeks universal algorithms	Acknowledges diversity but looks for common principles	Intelligence as multiple, incommensurable capacities
Benchmarks	Aggregates across domains (ARC-AGI, BIG-bench)	Domain-specific scores but with overall rankings	Separate evaluation per domain; refuses aggregation
Comparison Claims	"X is more intelligent than Y"; cross-species/-system rankings	Qualified comparisons ("more intelligent at task T")	Rejects cross-system comparison as category error
Scaling Laws	Evidence of AGI convergence	Sees progress but notes gaps	Benchmark-specific optimization
Failure Attribution	"Engineering problems"; "needs more scale/data"	Some architectural, some fundamental	"Fundamental mismatches"; qualitative differences
AI Risk Focus	AGI; univ. alignment principles	Both x-risk and near-term risks	Context-specific, near-term
Architecture Pref.	Unified, general-purpose models	Modular systems	Specialized domain/task models
Success Criteria	Domain-general competence	Task-specificity, some transfer	Specific cognitive phenomena

Table 1: Diagnostic Rubric for Identifying Intelligence Positions

Different Assumptions of Intelligence Assumptions

Before presenting our classification, we want to disclaim two points. First, by claiming that different researchers have different underlying assumptions of intelligence, we do not intend to imply that these necessarily have been *articulated*. We posit that assumptions are revealed through *methodological* choices and interpretation of empirical research, and will point below to these exact differences. Second, as mentioned above, by asserting that different views of intelligence exist, we do not intend to imply that these exist only under hard, declarative boundaries. Agreement with either position need not involve a wholesale endorsement of every aspect. We briefly develop an intermediate position below.

Intelligence Realism

Intelligence realism makes three distinct but related claims: (1) Ontological: There exists a single, universal computational process that constitutes intelligence across all systems; (2) Epistemological: This universal process is discoverable through scientific investigation; (3) Methodological: Different intelligent systems can be meaningfully ranked according to how well they approximate this universal process. Here, we detail its underlying assumptions.[1]

Algorithmic Universality The core ontological premise of intelligence realism is that intelligence follows universal and discoverable mathematical principles regardless of its instantiation. Epistemologically, this translates to the possibility of intelligence being described and tested for in abstract terms. One example of this perspective is Marcus Hutter's AIXI model (Hutter 2003), which is a theoretical framework that defines (artificial general) intelligence as the ability to achieve goals in different environments. Hutter's AIXI formulation assumes that optimal intelligence can be expressed as a single optimization function across all environments, presupposing that different environments are commensurable and that a single policy exists that maximizes value across all contexts.

This implies that realists interpret scaling laws as evidence for algorithmic universality: if larger models consistently improve across diverse tasks, this suggests convergence toward a universal intelligence algorithm. The smooth power-law relationships observed across language modeling, reasoning, and multimodal tasks support the realist claim that a single optimization process underlies all intelligent behavior.

Single Optimality Closely tied in with the algorithmic universality assumption, secondly, intelligence realism ontologically asserts the existence of an *optimal* intelligence algorithm. For this, consider intelligence as a universal algorithm that every instantiation (e.g., dolphin, human, or AI) can approximate.[2] If every instantiation of this algorithm is different, then every instantiation of this algorithm is closer or further away from *the* universal intelligence algorithm. It follows that there exists a ranking of every organism of how closely this algorithm is instantiated. We want to note that single optimality does not posit that all systems can be linearly ranked, but that there exists an optimal mapping from computational resources to performance across environments. Different systems might be optimal under different resource constraints.

One way this position is methodologically reflected in the literature is again the AIXI work (Hutter 2003). Here, expected rewards are counted over a specific time. The uni-

[1] We use "Intelligence Realism" to denote a position that combines ontological realism (intelligence exists as a real phenomenon) with monism (it constitutes a single, universal capacity). While these are technically separable commitments (some would say: orthogonal), with this notion, we posit that they travel together in the AI research context we analyze: Claims about universal intelligence algorithms typically presuppose their reality as computational processes.

[2] "Instantiation" refers to a particular physical or computational system that implements (more or less successfully) the abstract optimization function, much as different computers can instantiate the same sorting algorithm with varying efficiency.

versal "best" algorithms have the highest number of expected rewards for a specific time span, while instantiations of a "worse" algorithm will count less. Also, Schmidhuber's work on an "optimal ordered problem solver" (Schmidhuber 2004), a "general and in a certain sense time-optimal way of solving one problem after another," suggests that there is an optimal way of solving problems and that this property can be discovered.

Intelligence Variance Through Implementation Following from the above, realism also asserts that different manifestations of intelligence are viewed as *varying implementations* or approximations of the universal algorithmic core.[3] To be precise, realists do not assume identical algorithms, but rather that all intelligent behavior optimizes the same abstract objective function, as for example something like "maximize expected utility given computational constraints" (Legg and Hutter 2007; Hutter 2003). Different implementations (neural networks, biological brains) approximate this optimum differently, but the underlying optimization target is universal. This perspective is exemplified by comparative studies that attempt to measure intelligence across species using information-theoretic or computational metrics.

Consider the work of Commons and Ross (2008), which developed cross-species intelligence metrics based on task-solving efficiency and adaptive behavior. Their research suggests that despite significant differences in neural architecture, different species might be understood as instantiating similar underlying computational principles. A chimpanzee solving a tool-use problem, a corvid caching food, and an AI system playing chess could potentially be analyzed through a common algorithmic lens.

Reducibility Any advanced form of intelligence can be reduced to a combination of constitutive features. By doing so, it could reveal fundamental principles applicable across all intelligent systems. This reductionist approach finds methodological expression in efforts to create comprehensive intelligence benchmarks across different specific domains.

For example, Goertzel (2014) presents an attempt to develop tests that measure intelligence independent of a specific implementation. The most comprehensive effort in this direction is Hernández-Orallo (2017), which develops formal frameworks for evaluating diverse intelligent systems (from biological organisms to AI) using task-based measures that claim to transcend specific implementations. Similarly, Chollet (2019) proposes the Abstraction and Reasoning Corpus (ARC) as a measure of general fluid intelligence focused on skill-acquisition efficiency rather than task-specific performance. By designing tasks that require flexible problem-solving across diverse domains, researchers sought to create a "universal intelligence test" that could compare cognitive capabilities across different systems that transcend domain-specific knowledge.

[3]"Algorithmic core" is the abstract computational procedure that is independent of physical substrate and defines the "optimal" information processing for a certain goal.

Commensurability The final key assumption following the assumptions on variation through implementation and reducibility is that all forms of intelligence can, after all, be meaningfully compared along a single dimension or a small set of unified dimensions. This perspective underpins many intelligence tests, both natural and artificial. Without commensurability, the realist's claim that there exists an optimal intelligence algorithm becomes meaningless: if different forms of intelligence cannot be compared, then the notion of optimality loses its foundation.

As an example, Legg and Hutter (2007) argue that not only artificial intelligence but also intelligence across different species can be formalized as "an agent's ability to achieve goals in a wide range of environments." More specifically, their formal definition of universal intelligence Υ of an agent π, $\Upsilon(\pi) = \sum_{\mu \in E} w_\mu V_\mu^\pi$ encodes the realist assumption of commensurability by assigning weights w_μ to different environments μ of the space of all environments E. This implies that these environments can be meaningfully compared and aggregated into a single intelligence measure.[4]

Intelligence Pluralism

Intelligence pluralism denies all three realist claims, arguing instead that: (1) Ontologically, intelligence consists of multiple, incommensurable computational processes rather than varying implementations of a single process; (2) Epistemologically, these processes can only be understood *within* their specific ecological and evolutionary contexts because context is constitutive, not merely obscuring; (3) Methodologically, cross-system comparisons of "intelligence" are therefore either impossible or meaningless because there is no universal standard against which to measure.[5]

Algorithmic Diversity Ontologically, pluralism sees intelligence as inherently tied to the specific context, environment, and evolutionary history in which it develops. While realism assumes that there is one universal algorithm across species that intelligence can be represented as, pluralism assumes that there are many different algorithms across species and maybe even within species.

One such example is the navigational system of desert ants. Desert ants have evolved a path integration system that

[4]Historically, in human intelligence, this was also expressed as the g-factor (Eid et al. 2017), or the IQ (Terman 1916). The debate over whether g represents a real cognitive capacity or merely a useful statistical artifact mirrors the realism-pluralism distinction we develop here. Realists treat g as reflecting a unified cognitive capacity (Jensen 1998), while pluralists see it as an instrumental construct that may not correspond to any natural kind (see also Borsboom (2005) and Vessonen (2019)).

[5]These pluralist commitments connect to broader traditions in cognitive science. Gardner's theory of multiple intelligences (Gardner 1983), Gigerenzer's work on ecological rationality (Gigerenzer and Goldstein 1996), and enactivist approaches emphasizing embodied, embedded cognition (Wilson 2002) all challenge the notion of intelligence as a unified, context-independent capacity. These frameworks differ in details, yet share the core pluralist insight that intelligent behavior emerges from specific agent-environment couplings rather than universal computational principles.

is adapted to their featureless environment and calculates their return path using measurements of distance and direction traveled. This strategy would be suboptimal in more complex terrains where agents can rely on landmark-based navigation (Wehner, Srinivasan et al. 2003). This challenges the realist assumption of optimality. The ant's path integration system would be catastrophically suboptimal in forest environments where landmark navigation dominates due to potential obstacles. Yet both systems are "optimal" within their contexts. This poses a dilemma for realists: either (1) there's no universal optimum, or (2) optimality is context-dependent, which undermines the universality claim.

Realists might posit that pluralists confuse implementation diversity with algorithmic diversity: Yes, ants and humans navigate differently, but both implement approximate solutions to the same abstract problem: optimal path planning under uncertainty. The diversity is in implementation, not in the underlying computational problem.

Pluralists reject this move as an ontological error. The realist position assumes that "path planning" exists as a natural kind, i.e., as a problem discoverable through formal analysis that exists independently of any particular agent (see Magnus (2012); Boyle (2024)). Pluralists see this as an analyst's construction retrospectively imposed on diverse behaviors. Ants are not solving "path planning" but following chemical gradients that happened to be evolutionarily successful; the "problem" only exists from an external perspective that imposes human-like goals onto non-human systems. For pluralists, path-finding in ants and humans does not yield different solutions to the same problem. Rather, they solve fundamentally different problems because they are different agents in different environments. Methodologically, this means we cannot identify the "same" problem across systems because the problem itself is constituted differently for each agent-environment coupling.

Multiple Equilibria Rather than a single optimal algorithm, pluralism posits that numerous valid cognitive architectures exist, optimized for different niches and challenges. For instance, different bird species have evolved distinct spatial memory systems: food-caching corvids like Clark's nutcrackers develop extraordinary spatial memory for storing thousands of seed locations, while other birds rely on entirely different navigational strategies (Sherry, Jacobs, and Gaulin 1992). In AI, this translates to recognition that different machine learning architectures might be optimally suited to different problem domains. A neural network excelling at image recognition may be fundamentally ill-suited to causal reasoning tasks, and so on.

This aligns with bounded rationality approaches that emphasize how cognitive architectures are optimized for specific computational constraints and environmental regularities rather than universal optimality (Simon 1990; Lieder and Griffiths 2020). Different resource constraints (time, memory, energy) lead to qualitatively different optimal strategies, undermining the notion of a single best algorithm.

It further implies that these different strategies, as we will detail below, cannot be ranked. As a first intuition, consider the navigational skills of migratory birds, the social skills of elephant herds, or the distributed intelligence of slime molds, of which each represents a cognitive strategy that defies ranking or comparison. We will also detail below what implications this has on the meaningfulness of the word "intelligence".

Emergence Pluralism posits that intelligence emerges from dynamic interactions between agents and environments, rather than from a single set of abstract principles that can be *universally* generalized in mathematical form.[6] It assumes that every form of behavior is a way of solving some problem for a specific type of agent in a specific type of environment. In contrast to realism, it posits that even abstract optimization targets embed specific assumptions about what counts as "utility" and which computational constraints matter. These assumptions are inevitably shaped by particular evolutionary histories and environmental pressures, making supposedly universal targets actually context-dependent.

Realists would posit that emergence does not preclude universal principles. Physics has universal laws despite emergent phenomena. Why should intelligence not have universal computational principles despite contextual emergence? Pluralists would counter that physical universals govern simple interactions that scale up to complexity. "Intelligence", however, may be ontologically different: intentional states, semantic content, and goal-directed behavior may be irreducibly tied to *specific forms* of embodiment and environmental embedding (see Penny (1995); Cangelosi et al. (2015)), not merely difficult to study in the abstract but constituted by these specificities. The analogy to physics assumes precisely what's in dispute: that intelligence involves universal building blocks.

Here, the obvious question is whether every species that has emerged over millions of years can be counted as intelligent. If intelligence means "cognitive strategies that enable successful environmental adaptation," then any species that has survived evolutionary pressures possesses intelligence. If pluralism rejects universal standards for intelligence and instead evaluates systems within their own contexts, then most surviving biological systems meet the criteria for intelligence within their respective niches.

It follows that under pluralism, the term "intelligence" loses discriminatory power. If every successful cognitive strategy counts as "intelligent," then intelligence becomes equivalent to "any cognitive adaptation that works". A concept that applies to everything explains nothing and becomes analytically vacuous. Can this then still be a helpful classification to look at, e.g., risks of AI under the pluralist lens? The implication is that *other* properties become more analytically useful, such as danger for other agents, efficiency given the environment and agent features, and innovativeness in comparison with other agents. Importantly, these criteria are all *relational* in the sense that they are not universal or general but only valid for a specific set of species. We

[6] Pluralists do not deny that intelligent behavior can be modeled mathematically: ant navigation and human planning can both be formalized. Rather, they deny that these diverse formalizations can be reduced to or unified under a single universal mathematical framework.

expand more on (emergent) LLM capabilities below under *Implications for AI Research: Interpretation*.

Irreducibility Pluralism further posits that intelligence as a whole cannot be reduced to abstract concepts but has to be viewed in its environmental and agent-based complexity. For example, some principles of human intelligence provide a scaffolding for understanding some principles of AI, but the complexity of the whole "intelligence" may be fundamentally irreducible to mathematical principles: only task-by-task can an algorithm be extracted from a human problem-solving attempt and then emulated in an artificial system.

Famously, octopuses demonstrate a distributed intelligence where a significant portion of their neural processing occurs in their arms, which is a very different cognitive architecture from centralized brain-based intelligence (Godfrey-Smith 2016). This challenges realist reducibility assumptions more fundamentally. If intelligence can be "distributed" across arms with semi-autonomous processing, what is the unit of analysis? The realist must either (1) deny that arm-based processing constitutes intelligence or (2) accept that intelligence does not reduce to centralized algorithms, which undermines the universal algorithm thesis. Ontologically, this suggests intelligence is not a property of algorithms per se but of specific agent-environment-body configurations. Methodologically, this means we cannot extract "the intelligence algorithm" from the octopus any more than we can extract "the wetness algorithm" from water; it's an emergent property of the whole system that cannot be isolated from its constituent relations.

A realist might counter that this misidentifies the relevant unit of analysis: the unit is the optimization process, not the physical substrate, and that distributed processing just shows that optimal intelligence can be physically distributed while remaining computationally unified. However, pluralists might see this, again, as an ontological error: the realist move to abstract away from physical substrates assumes the very thing in question: that intelligence exists independently of its material and environmental instantiation. When the premise is to compare different systems, the specific agent-environment-body configuration is not an optional detail to be abstracted away but constitutive of what intelligence is in each case.

An example from the computational sciences is the field of computational cognitive science that methodologically tries to model the functions of the human brain *one algorithm at a time*. Examples are understanding epistemic language (Ying et al. 2025), virtual bargaining (Levine et al. 2024), logical reasoning (Olausson et al. 2023), online goal inference (Zhi-Xuan et al. 2020), or knowledge inference in lie production (Tan, Jara-Ettinger, and Berke 2024). In contrast to realist methodologies, these works are trying to figure out the exact algorithm for this one specific problem rather than trying to infer proxies for "intelligence" for a range of tasks from one model.

Incommensurability The last key assumption is that pluralism assumes that different forms of intelligence may be fundamentally "alien" to each other, prohibiting straightforward comparisons and understanding along universal dimensions. This follows from what we established above: if every system implements a range of specific, environmentally optimized solutions to its specific problem, a comparison to other systems would have to take into account this exact agent-environment pair. However, as soon as we compare it to the exact pair, we compare it to itself. As elaborated above, this also means that different systems cannot be ranked according to how "intelligent" they are and, with a high likelihood, can also not understand each other's "intelligence" as a whole, if this terminology is still useful at all. While realists would argue for an "optimization under constraints", pluralists would posit that this still assumes environments and constraints are commensurable enough to define meaningful optima. But if cognitive strategies are genuinely incommensurable, the entire optimization framework breaks down.

As a small example, researchers took a very long time until octopuses were ascribed something like "intelligent" behavior, and very likely this is still under an anthropomorphic lens. As an implication for neural language models, this suggests that these systems might develop forms of "intelligence" qualitatively different from human cognition. The models that excel at text generation, for instance, operate through fundamentally different mechanisms than human linguistic cognition. In this famous view, the term "understanding" would be inherently misplaced and overgeneralized when talking about a feature for artificial systems as they perform statistical transformations while lacking core aspects of human reasoning (see further (Mitchell 2021).

Intermediate Positions

The realism-pluralism distinction operates across multiple dimensions (ontological, epistemological, methodological), and researchers need not adopt consistent positions across all dimensions. For instance, one might hold *ontological realism* (believing intelligence reflects a universal computational process) while practicing *methodological pluralism* (using domain-specific evaluation frameworks because universal metrics are currently infeasible). Conversely, one might accept *ontological pluralism* (multiple incommensurable forms of intelligence) while employing *methodological reductionism* (attempting to decompose each form into constituent algorithms for practical analysis).

Implications for AI Research

Having outlined the realism-pluralism spectrum, we will now turn to the implications that different inherent positions on intelligence have on research methodology, the interpretation of results, and safety and governance.

Methodology

Methodological implications are the most direct and immediate consequences of an underlying intelligence position. Here, we look at model selection, benchmark design, and validation and success criteria.

Model Selection Even though both views agree on there being one brain in humans, the exact way in which this property can be realized in artificial models is the core of

the realist-pluralist debate. Realist assumptions favor unified architectures capable of general-purpose large language models, universal approximators, or architectures that can be scaled across diverse tasks. The popularity of transformer architectures partly reflects their apparent domain-generality.

Pluralist assumptions favor specialized architectures optimized for specific cognitive domains. Pluralists do not seek universal solutions but develop targeted approaches (as exemplified above): understanding epistemic language (Ying et al. 2025), virtual bargaining (Levine et al. 2024), logical reasoning (Olausson et al. 2023), online goal inference (Zhi-Xuan et al. 2020), or knowledge inference in lie production (Tan, Jara-Ettinger, and Berke 2024). Emulating a whole brain would be putting together all these specialized architectures in an efficient way (Griffiths, Chater, and Tenenbaum 2024).

Benchmark Design Realist assumptions naturally lead toward unified, domain-general benchmarks that assume different cognitive tasks tap into a common underlying capacity. BIG-Bench (Kazemi et al. 2025) or ARC-AGI (Chollet 2019) exemplify this approach. They assume that diverse domains across law, medicine, mathematics, or social reasoning (for BIG-Bench) or, again "intelligence" as a whole (for ARC-AGI) can be meaningfully aggregated to a single measure and thereby implicitly assume the domains share sufficient commonality to warrant aggregated scoring. However, multi-task benchmarks also show the spectrum nature of our framework. They do acknowledge domain distinctions (pluralist view) and yet still aggregate performance into unified rankings that assume commensurable underlying capacities (realist view), which reflects the intermediate positions we have touched upon above.

Pluralist methodology, by contrast, demands task-specific evaluations that respect the ecological context in which "intelligence" arises and operates. Rather than seeking a universal metric, pluralists insist on separate evaluation frameworks for each cognitive domain, designed around the specific environmental pressures and adaptive challenges that shaped those capacities. Examples of this specialization assumption are cognitive benchmarks like ConceptARC for abstract concept understanding (Moskvichev, Odouard, and Mitchell 2023), FANToM for Theory of Mind, (Kim et al. 2023), NormAd for adaptive norm understanding (Rao et al. 2024), or EWoK for world knowledge (Ivanova et al. 2024).

Validation and Success Criteria Realists judge success in terms of progress toward universal principles, i.e., theories that explain intelligence across species, domain-general competence, or across environments. Famous examples include AIXI (Hutter 2003), Hernández-Orallo and Dowe (2010), or Legg and Hutter (2007). They would further validate AI capabilities by demonstrating equivalence or superiority to biological intelligence on standardized tasks, as they assume commensurability between systems. Here, of course, the level of abstraction and the differentiation in architecture and functionality are of utmost importance. Some intermediate positions would argue that while architectural or fine-grained comparisons might be less meaningful, the system can functionally be compared or at some higher level of abstraction, as we detailed above.

Pluralist methodology judges success in terms of understanding *specific* cognitive phenomena within their ecological contexts. Rather than seeking universal explanations, pluralists aim for a rich understanding of how particular cognitive strategies solve specific adaptive challenges. For artificial systems, pluralists argue that these cannot be validated against biological intelligence because they solve fundamentally different computational problems through categorically different mechanisms. This, of course, creates big challenges for safety and governance: if we cannot assess how a machine is doing on metrics that make sense to us, how can we assess whether it will become dangerous or not? We will detail this below in the section on safety and governance.

Interpretation

Beyond shaping research methodology, realist and pluralist assumptions fundamentally alter how empirical evidence gets interpreted. The same data patterns become evidence for opposing theoretical positions, exemplifying what Kuhn (1962) called paradigm-dependent observation. This is not merely an abstract philosophical point. In practice, realist and pluralist researchers examine the same systems, read the same papers, and observe the same benchmark results, yet reach opposite conclusions about what these findings mean for AI capabilities and risks. The framework we developed provides a vocabulary for identifying when disagreements are empirical (solvable by more data) versus paradigmatic (requiring explicit philosophical examination).

Scaling Laws Two examples of opposite interpretations of the same data, scaling laws, are the realist "Sparks of Artificial General Intelligence" (Bubeck et al. 2023) and the pluralist "Why AI is Harder Than We Think" (Mitchell 2021).

Bubeck et al. (2023)'s central claim is that GPT-4 "could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system," which rests on the premise that intelligence constitutes a unified capacity manifesting across diverse cognitive domains. Multi-domain competence is interpreted as evidence for an underlying general intelligence progressing along a measurable continuum toward full AGI. The aggregated score reflects genuine cross-domain reasoning ability that approximates general intelligence.

The smooth scaling curves that show consistent improvement with increased parameters and training data are read as confirmation of algorithmic universality. Sudden capability jumps at critical scales are interpreted as genuine cognitive emergence, where quantitative parameter increases trigger qualitative leaps in reasoning ability (Wei et al. 2022).

Mitchell (2021) critiques these claims by proposing how some, as she claims, narrow advances in task-mastery are seen as the "first step" towards some more general goal, even though this progress might not lie on a continuum and an unexpected obstacle might always be in the way (Mitchell (2021) citing Dreyfus (2012)). For example, Deep Blue or GPT-2 have been claimed to be the first steps towards an "AI Revolution" (Aron 2016) or "general intelligence" (Alexander 2019). She supports this with an analogy by the engineer

and Hubert Dreyfus' brother, Stuart Dreyfus: "It was like claiming that the first monkey that climbed a tree was making progress towards landing on the moon" (Dreyfus 2012).

Under pluralist interpretation, scaling behaviors reflect sophisticated statistical pattern matching reaching various thresholds rather than genuine cognitive emergence. Multi-domain competence indicates that large datasets contain sufficient statistical regularities for pattern matching across diverse text types, not that the system possesses domain-general reasoning capabilities (see Kambhampati, Stechly, and Valmeekam (2025); Kambhampati et al. (2025)). The "emergence" of new capabilities represents accumulated statistical sophistication crossing perceptual thresholds, not the development of genuine understanding (Schaeffer, Miranda, and Koyejo 2023). Lastly, domain-specific performance ceilings reflect the inherent limitations of statistical approaches to cognitive tasks requiring genuine understanding (e.g., Isbilen and Christiansen (2022)).

Emergent Capabilities This interpretive difference extends to the very concept of "emergent abilities" in LLMs, which has become a focal point of the realism-pluralism debate. What realists interpret as discontinuous capability emergence, i.e., the sudden appearance of capabilities like arithmetic, theory of mind, or complex reasoning at scale, pluralists challenge as an artifact of measurement assumptions. First, apparent discontinuities may result from human-designed benchmarks with arbitrary pass/fail thresholds that make smooth capability development appear sudden (Schaeffer, Miranda, and Koyejo 2023). The "emergence" lives in the metrics, not the underlying process.

Second, and more fundamentally, pluralists question whether LLMs and humans develop "the same" capabilities at all. When an LLM produces arithmetic outputs, is it "doing arithmetic" in any sense comparable to human calculation, or performing statistical pattern matching that we retrospectively label as "arithmetic" because outputs coincide? The realist assumes "arithmetic" is a natural computational category that both humans and LLMs instantiate; the pluralist sees this as imposing human categories onto alien processes. This connects to benchmark selection: we evaluate LLMs on human-centric tasks and interpret their success as evidence of universal intelligence, assuming our capabilities represent the right targets for measuring intelligence. Realists interpret scaling laws as evidence of convergence toward universal intelligence; pluralists see them as evidence that we've successfully optimized for our own benchmarks.

Failure Analysis Realists frame observed deficits as architectural flaws or insufficient size, both of which can be solved by engineering and what some have termed "hardware fixes" (Musser 2018). When GPT-4 exhibits limitations in long-term planning, temporal reasoning, or causal understanding, Bubeck et al. (2023) characterize these as "missing components" amenable to engineering solutions through "further training." This treats general intelligence as achievable through scaling solutions rather than inherent ecological and qualitative advances.

Pluralists view such failures as revealing qualitative differences between human and artificial cognition that resist technical remediation rather than being incidental. In particular, they reveal fundamental architectural mismatches between statistical pattern matching and the contextual, embodied cognition that characterizes biological intelligence. For instance, they argue that large language models still struggle with common-sense physical reasoning (Ivanova et al. 2024), exhibit systematic failures in Theory of Mind (Kim et al. 2023), or do vision processing in an inherently different, non-human way (Bowers et al. 2023). They argue that systematic failures indicate fundamental limits, suggesting that human-like cognition requires qualitatively different architectures rather than incremental improvements to existing systems (e.g., Diaz and Madaio (2024); Griffiths, Chater, and Tenenbaum (2024))

Alignment

These interpretive differences have direct consequences for AI safety and governance.

Risk Assessment Realist assumptions naturally lead to concerns about superintelligent agents pursuing unified goals with domain-general capabilities. If intelligence constitutes a universal algorithm that can be optimized across all environments, and we see that some models are improving on one axis, then sufficiently advanced AI systems will surpass humans on that general intelligence scale. Systems will necessarily develop coherent goal structures and general competencies that enable them to pursue objectives across diverse domains (Bostrom 2014). Sufficiently advanced AI systems would develop universal instrumental goals, making control increasingly difficult as capabilities scale (Amodei et al. 2016).

This realist framework directly informs major alignment research programs. For instance, the Constitutional AI framework assumes harmful behaviors emerge from a single, trainable value system that can be shaped through universal principles (Bai et al. 2022). Similarly, scalable oversight presupposes that alignment properties transfer across capability levels, i.e., if we can align a weaker system, we can use it to align stronger systems because they implement the same underlying optimization process (Zeng et al. 2025; Bowman et al. 2022).

The realist assumption of algorithmic universality also underlies concerns about mesa-optimization and deceptive alignment (Hubinger et al. 2019). If intelligence follows universal principles, then sufficiently capable systems will inevitably develop internal optimization processes that may not align with their training objectives, leading to systematic deception during training that only manifests during deployment.

Pluralist assumptions generate fundamentally different threat models focused on diverse system behaviors emerging from the interaction of specialized AI systems operating in different contexts. The core premise that pluralists argue for is that the currently used psychometric tests investigate the wrong tasks while failing to capture crucial aspects of intelligence, such as meta-learning, causal reasoning, innate priors, and robust generalization, all considered essential for superintelligence (Raji et al. 2021; Mitchell 2021).

This leads to different risk priorities. Since current AI systems already demonstrate "intelligence", i.e., high sophistication, *within* specific domains like algorithmic bias in criminal justice, labor displacement (Gebru and Torres 2024), or misinformation (Bender 2024), immediate societal impacts in these domains might take precedence over *hypothetical* superintelligence scenarios (Moorosi, Sefala, and Luccioni 2023).

However, as touched upon above, the boundaries are not clear-cut, and there might be many researchers and research programs that acknowledge both superintelligence risk and societal risk (e.g., see Hendrycks, Mazeika, and Woodside (2023)).

Alignment Approaches Realists often seek universal principles to control or align systems and favor solutions like reward models that capture human preferences across all domains (Russell 2019; Zhi-Xuan et al. 2024a), interpretability techniques that reveal the universal optimization process underlying AI behavior (Elhage et al. 2022), and robustness guarantees that hold regardless of deployment context (Zhao et al. 2024; Sanfiz and Akrout 2021). Some of these attempts, however, face challenges that suggest the validity of a more pluralist view, including issues of context dependence (Millière 2023), frame problems (Shanahan 2004; Peterson 2025), and specification challenges (Zhi-Xuan et al. 2024a). For example, impact measures like Attainable Utility Preservation (Turner, Hadfield-Menell, and Tadepalli 2020) struggle to define impacts without reference to specific contexts and value systems. The persistent challenge of reward hacking across different alignment approaches further demonstrates how safety mechanisms designed under universalist assumptions often fail when confronted with the rich complexity of real-world environments (Amodei et al. 2016).

Pluralist priorities emphasize context-specific alignment: developing evaluation frameworks that assess AI behavior within specific ecological niches (Enevoldsen et al. 2023; Martin et al. 2019), specific cognitive functions (Zhi-Xuan et al. 2024b; Olausson et al. 2023), and specific human-AI collaboration contexts rather than general-purpose intelligence (Collins et al. 2024; Zhi-Xuan et al. 2024b; Oldenburg and Zhi-Xuan 2024). This includes research on cultural value diversity in AI systems (Rao et al. 2024; Li et al. 2024a,b), work on developing domain-specific safety measures for AI deployment in healthcare (Aggarwal et al. 2023), autonomous driving (Wäschle et al. 2022), education (Clark et al. 2025), etc., and investigation of how different AI architectures might be optimally suited to different human social contexts (Tomašev et al. 2020; Abbass 2019).

Governance Implications These different approaches lead to concrete policy disagreements. Realists tend to support capability-based regulation focusing on general AI capabilities regardless of application, and international coordination on universal AI safety standards (Hendrycks, Mazeika, and Woodside 2023). This perspective informs proposals for global AI governance institutions and capability-based licensing schemes.

Pluralists advocate for application-specific regulation that varies by domain and context, emphasizing democratic participation in AI governance and regulatory frameworks that account for the diversity of human values and social contexts (Bogiatzis-Gibbons 2024). The EU AI Act exemplifies this approach by categorizing risks based on deployment context rather than general capability level (Veale and Zuiderveen Borgesius 2021).

Conclusion

We have argued that contemporary AI research operates along a spectrum between two fundamentally different conceptions of intelligence that remain largely implicit: Intelligence Realism and Intelligence Pluralism. We showed how they fundamentally shape the field's methodology, interpretation, and risk assessment. Making these underlying assumptions explicit serves several purposes.

Methodologically, recognizing the realism-pluralism spectrum helps researchers understand why different approaches to benchmark design, model selection, and validation criteria persist. These are not merely engineering choices but reflect deep commitments about the nature and existence of "intelligence". Researchers favoring a domain-specific architecture and task-specific, cognitive benchmarks probably agree more with the pluralist framing than those favoring general models and aggregate benchmarks. More importantly, they are operating from different ontological starting points.

Interpretively, the framework clarifies why identical empirical evidence generates opposite conclusions. When realists and pluralists examine GPT-4's performance, they are not disagreeing about the data but about what "intelligence" is and how it can be recognized. This explains why scaling law debates, failure analysis, and capability assessments remain unresolved despite abundant empirical work. The disagreements are paradigmatic, not merely empirical.

For AI risk research, these different conceptions generate fundamentally different risk models and governance approaches. Realists focus on superintelligence scenarios and universal alignment principles; pluralists emphasize distributed, context-specific, and currently existing risks and tailored safety measures. Both perspectives identify risks, but they prioritize different threats and solutions. Making these commitments explicit enables more productive debates: rather than arguing past each other about "the" AI risk, stakeholders can identify which conception of intelligence underlies their concerns and whether disagreements are empirical or philosophical.

We do not claim to resolve the realism-pluralism debate that would require extensive empirical and philosophical work beyond this paper's scope. Rather, our contribution is to make explicit the implicit assumptions that structure contemporary AI research, providing a vocabulary for clearer scientific and policy discourse. As AI systems become more capable and consequential, understanding these foundational disagreements becomes increasingly urgent. A research community that explicitly engages with its philosophical commitments is better positioned to navigate the challenges ahead.

Acknowledgements

We thank members and visitors of the Hong Kong AI & Humanity lab for their fruitful comments and feedback on earlier versions of this paper, especially Rachel Sterken, Seth Lazar, Nate Sharadin, Herman Cappelen, and Kate Vredenburgh. We also thank the five anonymous reviewers for their constructive feedback that helped to shape the paper.

References

Abbass, H. A. 2019. Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, 11(2): 159–171.

Aggarwal, A.; Tam, C. C.; Wu, D.; Li, X.; and Qiao, S. 2023. Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. *Journal of medical Internet research*, 25: e40789.

Alexander, S. 2019. GPT-2 As Step Toward General Intelligence. *Slate Star Codex*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Aron, J. 2016. AI landmark as Googlebot beats top human at ancient game of Go. *New Scientist*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bender, E. M. 2024. Resisting dehumanization in the age of "AI". *Current Directions in Psychological Science*, 33(2): 114–120.

Bengio, Y.; Mindermann, S.; Privitera, D.; Besiroglu, T.; Bommasani, R.; Casper, S.; Choi, Y.; Fox, P.; Garfinkel, B.; Goldfarb, D.; et al. 2025. International AI Safety Report. *arXiv preprint arXiv:2501.17805*.

Bogiatzis-Gibbons, D. J. 2024. Beyond Individual Accountability:(Re-) Asserting Democratic Control of AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 74–84.

Borsboom, D. 2005. *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Bostrom, N. 2014. Superintelligence: Paths, dangers, strategies.

Bowers, J. S.; Malhotra, G.; Dujmović, M.; Montero, M. L.; Tsvetkov, C.; Biscione, V.; Puebla, G.; Adolfi, F.; Hummel, J. E.; Heaton, R. F.; et al. 2023. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 46: e385.

Bowman, S. R.; Hyun, J.; Perez, E.; Chen, E.; Pettit, C.; Heiner, S.; Lukošiūtė, K.; Askell, A.; Jones, A.; Chen, A.; et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Boyle, A. 2024. Disagreement & classification in comparative cognitive science. *Noûs*, 58(3): 825–847.

Bubeck, S.; Chadrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Cangelosi, A.; Bongard, J.; Fischer, M. H.; and Nolfi, S. 2015. Embodied intelligence. *Springer handbook of computational intelligence*, 697–714.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Clark, H.-B.; Benton, L.; Searle, E.; Dowland, M.; Gregory, M.; Gayne, W.; and Roberts, J. 2025. Building Effective Safety Guardrails in AI Education Tools. In *International Conference on Artificial Intelligence in Education*, 129–136. Springer.

Collins, K. M.; Sucholutsky, I.; Bhatt, U.; Chandra, K.; Wong, L.; Lee, M.; Zhang, C. E.; Zhi-Xuan, T.; Ho, M.; Mansinghka, V.; et al. 2024. Building machines that learn and think with people. *Nature human behaviour*, 8(10): 1851–1863.

Commons, M. L.; and Ross, S. N. 2008. Toward a cross-species measure of general intelligence. *World Futures*, 64(5-7): 383–398.

Diaz, F.; and Madaio, M. 2024. Scaling laws do not scale. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 341–357.

Dreyfus, H. L. 2012. A history of first step fallacies. *Minds and Machines*, 22(2): 87–99.

Eid, M.; Geiser, C.; Koch, T.; and Heene, M. 2017. Anomalous results in G-factor models: Explanations and alternatives. *Psychological methods*, 22(3): 541.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Enevoldsen, K.; Hansen, L.; Nielsen, D. S.; Egebæk, R. A.; Holm, S. V.; Nielsen, M. C.; Bernstorff, M.; Larsen, R.; Jørgensen, P. B.; Højmark-Bertelsen, M.; et al. 2023. Danish foundation models. *arXiv preprint arXiv:2311.07264*.

Gardner, H. 1983. *Frames of mind: The theory of multiple intelligences*.

Gebru, T.; and Torres, É. P. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*.

Gigerenzer, G.; and Goldstein, D. G. 1996. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4): 650.

Godfrey-Smith, P. 2016. *Other minds: The octopus, the sea, and the deep origins of consciousness*. Farrar, Straus and Giroux.

Goertzel, B. 2014. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1): 1.

Griffiths, T. L.; Chater, N.; and Tenenbaum, J. B. 2024. *Bayesian models of cognition: Reverse engineering the mind*. MIT Press.

Hanson, N. R. 1958. The logic of discovery. *The Journal of Philosophy*, 55(25): 1073–1089.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An overview of catastrophic AI risks. *arXiv preprint arXiv:2306.12001*.

Hernández-Orallo, J. 2017. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press.

Hernández-Orallo, J.; and Dowe, D. L. 2010. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18): 1508–1539.

Hubinger, E.; van Merwijk, C.; Mikulik, V.; Skalse, J.; and Garrabrant, S. 2019. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

Hutter, M. 2003. A gentle introduction to the universal algorithmic agent AIXI.

Isbilen, E. S.; and Christiansen, M. H. 2022. Statistical learning of language: A meta-analysis into 25 years of research. *Cognitive Science*, 46(9): e13198.

Ivanova, A. A.; Sathe, A.; Lipkin, B.; Kumar, U.; Radkani, S.; Clark, T. H.; Kauf, C.; Hu, J.; Pramod, R.; Grand, G.; et al. 2024. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Jensen, A. R. 1998. The factor. *Westport, CT: Prager*.

Johnson, S. G.; Karimi, A.-H.; Bengio, Y.; Chater, N.; Gerstenberg, T.; Larson, K.; Levine, S.; Mitchell, M.; Rahwan, I.; Schölkopf, B.; et al. 2024. Imagining and building wise machines: The centrality of AI metacognition. *arXiv preprint arXiv:2411.02478*.

Kambhampati, S.; Stechly, K.; and Valmeekam, K. 2025. (How) Do reasoning models reason? *Annals of the New York Academy of Sciences*, 1547(1): 33–40.

Kambhampati, S.; Stechly, K.; Valmeekam, K.; Saldyt, L.; Bhambri, S.; Palod, V.; Gundawar, A.; Samineni, S. R.; Kalwar, D.; and Biswas, U. 2025. Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces! *arXiv preprint arXiv:2504.09762*.

Kazemi, M.; Fatemi, B.; Bansal, H.; Palowitch, J.; Anastasiou, C.; Mehta, S. V.; Jain, L. K.; Aglietti, V.; Jindal, D.; Chen, P.; et al. 2025. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R. L.; Kim, G.; Choi, Y.; and Sap, M. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

Kuhn, T. 1962. *The nature and necessity of scientific revolutions*. na.

Legg, S.; and Hutter, M. 2007. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17(4): 391–444.

Levine, S.; Kleiman-Weiner, M.; Chater, N.; Cushman, F.; and Tenenbaum, J. B. 2024. When rules are over-ruled: Virtual bargaining as a contractualist method of moral judgment. *Cognition*, 250: 105790.

Li, C.; Chen, M.; Wang, J.; Sitaram, S.; and Xie, X. 2024a. Culturellm: Incorporating cultural differences into large language models. *Advances in Neural Information Processing Systems*, 37: 84799–84838.

Li, C.; Teney, D.; Yang, L.; Wen, Q.; Xie, X.; and Wang, J. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Advances in Neural Information Processing Systems*, 37: 65183–65216.

Lieder, F.; and Griffiths, T. L. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43: e1.

Magnus, P. D. 2012. *Scientific enquiry and natural kinds: From planets to mallards*. Springer.

Martin, L.; Muller, B.; Suárez, P. J. O.; Dupont, Y.; Romary, L.; de La Clergerie, É. V.; Seddah, D.; and Sagot, B. 2019. CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894*.

Millière, R. 2023. The alignment problem in context. *arXiv preprint arXiv:2311.02147*.

Mitchell, M. 2021. Why AI is harder than we think. *arXiv preprint arXiv:2104.12871*.

Mitchell, M. 2024. Debates on the nature of artificial general intelligence.

Moorosi, N.; Sefala, R.; and Luccioni, S. 2023. AI for whom? Shedding critical light on AI for social good. In *NeurIPS 2023 Computational Sustainability: Promises and Pitfalls from Theory to Deployment*.

Moskvichev, A.; Odouard, V. V.; and Mitchell, M. 2023. The conceptarc benchmark: Evaluating understanding and generalization in the arc domain. *arXiv preprint arXiv:2305.07141*.

Musser, G. 2018. Job one for quantum computers: Boost artificial intelligence. *Quanta Magazine*.

Olausson, T. X.; Gu, A.; Lipkin, B.; Zhang, C. E.; Solar-Lezama, A.; Tenenbaum, J. B.; and Levy, R. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. *arXiv preprint arXiv:2310.15164*.

Oldenburg, N.; and Zhi-Xuan, T. 2024. Learning and sustaining shared normative systems via bayesian rule induction in markov games. *arXiv preprint arXiv:2402.13399*.

Penny, S. 1995. Embodied Mind: Cognitive Science and Human Experience by Francisco J. Varela, Evan Thompson, Eleanor Rosch. *Leonardo*, 28(4): 337–338.

Peterson, J. 2025. Context Sensitive Frames and AI Alignment. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, 1251–1254. IEEE.

Phan, L.; Gatti, A.; Han, Z.; Li, N.; Hu, J.; Zhang, H.; Shi, S.; Choi, M.; Agrawal, A.; Chopra, A.; et al. 2025. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*.

Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Rao, A.; Yerukola, A.; Shah, V.; Reinecke, K.; and Sap, M. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*.

Russell, S. 2019. *Human compatible: AI and the problem of control*. Penguin Uk.

Sanfiz, A. J.; and Akrout, M. 2021. Benchmarking the accuracy and robustness of feedback alignment algorithms. *arXiv preprint arXiv:2108.13446*.

Schaeffer, R.; Miranda, B.; and Koyejo, S. 2023. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36: 55565–55581.

Schmidhuber, J. 2004. Optimal ordered problem solver. *Machine Learning*, 54(3): 211–254.

Shanahan, M. 2004. The frame problem.

Sherry, D. F.; Jacobs, L. F.; and Gaulin, S. J. 1992. Spatial memory and adaptive specialization of the hippocampus. *Trends in neurosciences*, 15(8): 298–303.

Simon, H. A. 1990. Bounded rationality. In *Utility and probability*, 15–18. Springer.

Tan, Z. Y.; Jara-Ettinger, J.; and Berke, M. 2024. Reasoning about knowledge in lie production. Proceedings of the Annual Meeting of the Cognitive Science Society.

Terman, L. M. 1916. *The measurement of intelligence*, volume 191. Houghton Mifflin Company Boston.

Tomašev, N.; Cornebise, J.; Hutter, F.; Mohamed, S.; Picciariello, A.; Connelly, B.; Belgrave, D. C.; Ezer, D.; Haert, F. C. v. d.; Mugisha, F.; et al. 2020. AI for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1): 2468.

Turner, A. M.; Hadfield-Menell, D.; and Tadepalli, P. 2020. Conservative agency via attainable utility preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 385–391.

Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4): 97–112.

Vessonen, E. S. M. 2019. *Representing and Constructing. Psychometrics from the perspectives of measurement theory and concept formation.* Ph.D. thesis.

Wäschle, M.; Thaler, F.; Berres, A.; Pölzlbauer, F.; and Albers, A. 2022. A review on AI Safety in highly automated driving. *Frontiers in artificial intelligence*, 5: 952773.

Wehner, R.; Srinivasan, M. V.; et al. 2003. Path integration in insects. *The neurobiology of spatial behaviour*, 9–30.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wilson, M. 2002. Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4): 625–636.

Ying, L.; Zhi-Xuan, T.; Wong, L.; Mansinghka, V.; and Tenenbaum, J. B. 2025. Understanding epistemic language with a language-augmented bayesian theory of mind. *Transactions of the Association for Computational Linguistics*, 13: 613–637.

Zeng, Y.; Zhao, F.; Wang, Y.; Lu, E.; Yang, Y.; Wang, L.; Liu, C.; Liang, Y.; Zhao, D.; Han, B.; et al. 2025. Redefining Superalignment: From Weak-to-Strong Alignment to Human-AI Co-Alignment to Sustainable Symbiotic Society. *arXiv preprint arXiv:2504.17404*.

Zhao, Y.; Yan, L.; Sun, W.; Xing, G.; Wang, S.; Meng, C.; Cheng, Z.; Ren, Z.; and Yin, D. 2024. Improving the robustness of large language models via consistency alignment. *arXiv preprint arXiv:2403.14221*.

Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2024a. Beyond preferences in ai alignment. *Philosophical Studies*, 1–51.

Zhi-Xuan, T.; Mann, J.; Silver, T.; Tenenbaum, J.; and Mansinghka, V. 2020. Online bayesian goal inference for boundedly rational planning agents. *Advances in neural information processing systems*, 33: 19238–19250.

Zhi-Xuan, T.; Ying, L.; Mansinghka, V.; and Tenenbaum, J. B. 2024b. Pragmatic instruction following and goal assistance via cooperative language-guided inverse planning. *arXiv preprint arXiv:2402.17930*.