How Smart is Smart? A Critical Interdisciplinary Perspective on Artificial General Intelligence

Thomas Zoëga Ramsøy *

Posted Date: 21 March 2025

doi: 10.20944/preprints202503.1581.v1

Keywords: artificial general intelligence; large language models; psychometrics; philosophy of mind; emergence; AI benchmarks

 doi:10.20944/preprints202503.1581.v1

Article

How Smart is Smart? A Critical Interdisciplinary Perspective on Artificial General Intelligence

Thomas Zoëga Ramsøy

[1] Neurons Høje Taastrup Boulevard 33, 2650 Taastrup, Denmark; thomas@neuronsinc.com
[2] International Center for Applied Neuroscience, Trollesvej 6, 4581 Rørvig, Denmark
[3] Singularity University, 2831 Mission College Blvd, Santa Clara, California

Abstract: The concept of Artificial General Intelligence (AGI)—a machine capable of performing any intellectual task a human can—is both a central aspiration and a contested notion in AI research. Despite its prominence in scholarly and public discourse alike, AGI relies on unsettled definitions of intelligence and speculative assumptions about generalization. This paper critically examines AGI from multiple perspectives: conceptual theory, philosophy, psychometrics, and recent developments in large language models (LLMs). The foundations of AGI are undermined by the lack of consensus on what constitutes "intelligence" in both human and artificial contexts. Furthermore, I explore how AGI systems may excel at benchmarks by optimizing for performance rather than demonstrating genuine understanding—akin to the "simulation without comprehension" phenomenon described by Searle's Chinese Room argument. I also investigate the emergent behaviors reported in advanced AI models, assess whether these indicate genuine steps toward general intelligence or illusory artifacts, and discuss how introspective features in LLMs might or might not constitute a move toward "self-awareness." By integrating insights from multiple disciplines, I propose a framework for reevaluating AGI that prioritizes scientific rigor, conceptual clarity, and ethical considerations. This analysis underscores the urgent need to distinguish mere test-passing behavior from true intelligence and to develop robust, psychometrically grounded benchmarks for AGI evaluation.

Keywords: artificial general intelligence; large language models; psychometrics; philosophy of mind; emergence; AI benchmarks

1. Introduction

Artificial General Intelligence (AGI) represents a cornerstone aspiration of artificial intelligence research, aiming to create machines that replicate or surpass human intellectual capabilities in diverse domains [1,2]. Despite progress in narrow AI systems that excel in specific tasks [3], the notion of AGI remains contentious. This paper critically examines AGI from multiple angles: conceptual foundations [4], philosophical implications [6], psychometric challenges [7], and recent debates on emergent properties in large language models.

The concept of intelligence itself is unsettled. Ongoing debates in psychology and neuroscience question whether intelligence is best understood as a single general capacity [8] or an ensemble of modular, domain-specific abilities [9]. These unresolved issues further complicate efforts to extend human intelligence frameworks to artificial systems [10]. In addition, many AGI benchmarks focus on task performance without adequately addressing whether success indicates genuine understanding or narrow optimization. Philosophical critiques such as Searle's Chinese Room [11] caution against conflating performance with comprehension, a concern amplified by the optimization-driven nature of modern AI [5,12].

Recent advances, including the widespread use of large language models (LLMs), have sparked discussions about possible emergent properties and limited forms of introspection. Whether these properties signify truly general intelligence or are merely artifacts of scale remains controversial.

 doi:10.20944/preprints202503.1581.v1

Finally, overclaims about AGI can pose societal risks, from eroding public trust to misdirecting research priorities [13,14].

This paper aims to provide a comprehensive assessment of the state of AGI research by weaving together insights from conceptual theory, philosophy, empirical AI research, and psychometrics. It highlights open questions and proposes directions for more rigorous and responsible development.

2. Conceptual Foundations

2.1. The Elusive Nature of Intelligence

Theories of human intelligence vary widely in their assumptions about its nature. Classical theories, such as Spearman's "g-factor," posit that intelligence is a general cognitive ability underlying performance across tasks [15]. In contrast, Howard Gardner's theory of multiple intelligences suggests that intelligence is domain-specific, comprising distinct types like linguistic, spatial, and interpersonal intelligence [16]. These divergent views illustrate the challenge of finding a universal definition that accommodates both general and specialized cognition.

Modern cognitive neuroscience raises further doubts about the unity of intelligence. Neural modularity research suggests cognitive functions are distributed across specialized brain regions [17]. Language processing, for instance, involves frontal and temporoparietal brain regions (now going beyond the classical Broca's and Wernicke's regions), while spatial reasoning critically depends on the parietal cortex as well as posterior medial temporal lobe regions [18].

This semi-modularity complicates any attempt to model intelligence as a single, uniform faculty and suggests that intelligence might emerge from the dynamic interplay of specialized subsystems. To make matters even more complicated, the brain and mind does not operate as a Fodorian strict modularism [38], but is littered with complex intertwining and network effects, and concepts such as degeneracy, redundancy, and pluripotentiality are better description of how we see the organization of the brain and mind today [37].

2.2. General Intelligence vs. Modularity

The existence of a dichotomy or fluent borders between general intelligence and modularity has direct implications for AGI. Many assume that creating a general-purpose AI is analogous to replicating a human-like "general intelligence." However, if human cognition is inherently semi-modular [17], such a goal may be conceptually flawed. Indeed, advanced AI systems that surpass human experts in narrowly defined tasks—such as playing chess [19] or generating human-like text [20]—often fail at broader generalization. These successes rely on specialized architectures and training datasets, lacking the breadth and adaptability of human intelligence [21]. Even so, the current messy re-evaluation of modularity and with new contents such as redundancy and pluripotentiality, any discussion or claim of AGI cannot be taken too lightly.

This gap between narrow brilliance and broad adaptability suggests that AGI may require not a single "master algorithm" but an integration of multiple more or less specialized functional modules [22]. Claims about AI reaching human-level general intelligence often neglect this complexity, conflating the mastery of one modality with true generality. The result is a persistent tension: while some continue pursuing a unifying approach to intelligence, evidence from both neuroscience and empirical AI achievements indicates a mosaic of specialized competencies.

2.3. Introspection in Large Language Models (LLMs)

A recent development in AI is the suggestion that large language models may exhibit rudimentary forms of *introspection*, akin to a system's ability to assess its own behavior. Studies have shown that when asked about their likely answers or confidence, LLMs can generate meta-statements that appear self-reflective [23]. Researchers have even found that one model can better predict its own future outputs than a second model trained solely on its past responses, suggesting some form of "internal representation" not trivially gleaned from external data [24].

 doi:10.20944/preprints202503.1581.v1

However, caution is warranted in labeling these behaviors as genuine introspection. LLMs remain pattern-matching entities trained on vast corpora of text, and their "self-reflective" statements may simply echo textual patterns relating to introspection. They lack a grounded, embodied self-model that human metacognition relies on. While introspection-like capabilities might enhance an AI's reliability (e.g., by allowing it to refuse tasks it deems it cannot do well), whether this represents progress toward true self-awareness or "general" introspective intelligence is debatable, even doubtful. Basically, introspection-like behavior can exist without there actually being introspection as in humans.

2.4. Emergent Properties: Genuine or Illusory?

Another hotly debated topic is the alleged emergence of new capabilities in AI systems once they exceed certain scale thresholds. Some observers have noted that large models demonstrate abilities—such as multi-step reasoning or complex analogies—not present in their smaller counterparts, framing these as "phase transitions" in AI [25]. Others argue these abrupt jumps may be artifacts of discrete evaluation metrics or test contamination.

For instance, a model might leap from 5% to 20% accuracy on a difficult test because it finally achieves partial competence, with a binary scoring metric making the jump look dramatic. Fine-grained analysis often reveals more incremental improvements. While true emergence is not impossible—biology itself shows emergent complexity—current evidence suggests that many claims of emergent "general intelligence" in LLMs may be overstated [26].

Practically, if emergent behaviors *do* occur, they pose both exciting and worrying prospects. On the one hand, scaling might uncover unexpected competencies. On the other, unpredictability complicates safety and alignment, as a system might spontaneously acquire abilities not anticipated by its creators. At the core of this, there is a substantial uncertainty and debate about whether emergentism exists even in biology, so whether AI models could in principle show emergent properties is likely even more debatable. However, even the possibility of emergent properties in AI models should warrant more research and debate. Thus, further research is needed to distinguish genuine emergent intelligence from measurement artifacts.

2.5. Conclusion of Conceptual Challenges

In sum, the unsettled nature of intelligence complicates the claims of AGI. Psychology and cognitive neuroscience suggests intelligence may be best understood as an ensemble of specialized components, but where these parts are not neatly defined biological units but rather the result of an ensemble of neural networks, where some of the same nodes in a given network can take on multiple roles. At the same time, philosophical and empirical work in AI indicates that large, seemingly comprehensive models still exhibit narrow optimization rather than broad cognition. Although LLM introspection and apparent emergent behaviors ignite optimism about bridging this gap, they remain subject to interpretative pitfalls. Moving forward, AGI researchers must be explicit about which aspects of intelligence they aim to replicate and wary of equating domain-specific breakthroughs with genuine generality.

3. Philosophical Critiques

The pursuit of Artificial General Intelligence (AGI) not only raises significant technical challenges but also invites profound philosophical scrutiny. Central to these debates is the question of whether an AI system—however advanced—can genuinely replicate the qualities of human understanding, consciousness, or intentionality. Philosophical critiques often focus on the distinction between the appearance of intelligence and the reality of understanding, questioning whether computational algorithms can ever transcend the mechanistic manipulation of symbols to achieve true semantic insight.

One of the most influential and enduring arguments in this field is John Searle's Chinese Room thought experiment. This critique sheds critical light on claims about the nature of intelligence and

 doi:10.20944/preprints202503.1581.v1

understanding in AI, directly challenging the proposition that computational systems, regardless of their apparent capabilities, can possess genuine understanding.

3.1. Searle's Chinese Room Argument

John Searle's Chinese Room argument [11] remains a foundational critique of the idea that computational systems—no matter how sophisticated—can truly "understand" in the human sense. It serves as a rebuttal to the claims of Strong AI (the view that a sufficiently advanced computer could achieve genuine understanding and mental states akin to humans). Searle introduces a thought experiment designed to show the limits of computational processes when it comes to replicating human-like semantic understanding.

In the Chinese Room scenario, a person who does not understand Chinese is placed in a room with instructions (a rulebook) for manipulating Chinese symbols. When given an input in Chinese, the person uses the instructions to produce an appropriate output in Chinese, without ever comprehending the language. To an external observer, the outputs might seem indistinguishable from those of a fluent Chinese speaker. However, Searle argues that, despite the functional performance displayed, there is no actual understanding of Chinese within the person or the system as a whole. The individual is merely following syntactic rules to manipulate symbols, without any grasp of their meaning.

The implications of this argument are far-reaching, especially in the context of modern machine learning systems such as Large Language Models (LLMs). These systems excel at generating contextually and syntactically appropriate outputs by processing vast amounts of data. Yet, as Searle's argument demonstrates, their ability to manipulate symbols does not necessarily entail an understanding of the semantics behind them. Machines, like the person in the Chinese Room, may exhibit superficial functionality that mimics intelligence while lacking the underlying comprehension that would qualify as true understanding.

The Chinese Room highlights the crucial distinction between *syntax* and *semantics*. Syntax refers to the rules for structuring and manipulating symbols, while semantics refers to the meaning and understanding associated with those symbols. Searle's critique can be used to emphasize that AI systems operate exclusively at the syntactic level, processing and generating symbols without any inherent understanding of their meaning. This fundamentally challenges the benchmarks used to assess AI progress, such as tasks that prioritize external performance alone.

Searle's thought experiment provokes essential questions: Is meaningful understanding necessary for intelligence? Or is functional performance sufficient for attributing intelligence to a system? While proponents of Strong AI might argue that indistinguishability in behavior between humans and machines (as suggested by the Turing Test) suffices as evidence of intelligence, Searle contends that the Chinese Room demonstrates why such functional tests are inadequate. Understanding, according to Searle, is not just about generating correct responses but entails intentionality—a quality that computational processes, rooted in synthetic manipulation, cannot achieve.

In the modern context, Searle's argument remains highly relevant. As contemporary AI systems produce results that increasingly resemble human outputs, critiques like the Chinese Room remind us to interrogate the philosophical underpinnings of these technologies and to carefully assess claims about their nature and capabilities. By emphasizing the limits of syntax-focused approaches, Searle's work urges ongoing reflection on the meaning of intelligence, understanding, and the potential distinctions between human cognition and artificial computation.

3.2. The Test-Passing Illusion

Benchmarks like the Turing Test [39] risk conflating *performance* with *intelligence*. A system that can fool human judges into believing it is human may still rely on shallow heuristics. Language models such as GPT-3 [20] and now later GPT4, as well as the more recent reasoning models can score impressively on many tasks, but do these scores denote genuine comprehension or sophisticated optimization?

 doi:10.20944/preprints202503.1581.v1

Here, it is vital to warn of a "test-passing illusion," wherein AI systems outperform humans on carefully curated tests yet fail at tasks requiring broad, flexible reasoning. For instance, a system might memorize patterns from its training set that appear to exhibit reasoning skills. When tested in slightly altered contexts, the facade of competence can crumble.

3.3. Anthropomorphism and Bias in Evaluation

Human evaluators often project human-like traits onto AI, a phenomenon known as anthropomorphism [27]. Fluent text generation or apparently introspective statements can lead observers to attribute understanding or self-awareness. This bias underlines the importance of rigorous, transparent benchmarks that distinguish truly cognitive processes from superficial mimicry.

For example, Nass and colleagues' early work highlighted that humans tend to apply social norms and human traits to computers, even in simple cases of interaction with automated systems [27]. This research forms part of a broader body of evidence indicating that even minimal social cues—such as a voice interface or responsiveness—can trigger anthropomorphism. Waytz, Cacioppo, and Epley (2010) further explored how anthropomorphic design choices in AI systems, such as providing a name, personality traits, or "emotions," strengthen these tendencies [44].

This tendency has considerable implications for emotional responses, trust, and behavior. For example, early on in the history of human-computer interaction, the "Eliza effect" demonstrated how users interacting with the early natural language processing chatbot ELIZA attributed emotional understanding and intentions to what was nothing more than a set of simple pattern-matching algorithms [40]. Despite ELIZA's simplicity, many participants reported feeling as though they were understood or cared for. Modern systems such as ChatGPT amplify this effect with their ability to fluently emulate conversational patterns, creating the illusion of empathy or moral reasoning.

Empirical studies have consistently found that human-like interactions increase trust and acceptance of AI, even in critical scenarios. Madhavan and Wiegmann (2007) have shown that anthropomorphic systems are trusted more than non-anthropomorphic ones, even after observable mistakes [41]. Similarly, de Visser et al. (2017) found that this increased trust can lead to over-reliance on AI systems, creating ethical concerns when such systems are used in high-stakes settings, such as autonomous vehicles, healthcare, or criminal justice [43].

From an ethical perspective, anthropomorphism can lead to misplaced trust in systems that lack the actual capabilities that human evaluators perceive. For example, if a chatbot is perceived as empathetic or morally capable based solely on its fluent text generation, real-world decisions might be inappropriately delegated to it. Such risks have already been documented in domains like mental health apps or virtual counseling tools, where perceived empathy from AI could mislead users into assuming correlations with therapeutic efficacy [42].

To address these challenges, rigorous public education and clear guidelines for AI developers are essential. Researchers must also develop benchmarks and evaluation methods to disentangle cognitive understanding from superficial simulation. Without such precautionary measures, anthropomorphic biases risk exacerbating ethical dilemmas with real-world consequences.

3.4. Implications for AGI Development

Philosophical critiques like Searle's Chinese Room, the test-passing illusion, and the anthropomorphizing bias underscore the need for evaluation methods that do more than measure surface behavior. As LLMs grow more capable at mimicking human dialogue, the risk of misjudging performance for intelligence increases. To move beyond performance-based illusions, AGI research must be better at digging into cognitive *mechanisms*, investigating whether AI systems conceptualize meaning or merely generate plausible outputs via pattern matching.

 doi:10.20944/preprints202503.1581.v1

4. Critique of AGI Benchmarks

4.1. Benchmark Limitations and Psychometric Gaps

AGI benchmarks, from the Turing Test to newer proposals like ARC-AGI, often lack robust psychometric underpinnings [28]. Validity, reliability, and standardization—the cornerstones of psychological testing—are seldom rigorously applied in AI. Some benchmarks may inadvertently measure memorization rather than reasoning or penalize methods that differ from human strategies.

Contamination of test data is another major issue. Large datasets can unintentionally include content from the very benchmarks meant to assess model capability [29,30]. When this happens, an AI's success might stem from prior exposure, inflating scores without reflecting real generalization. This poses a challenge, in the sense that AI model testing hinges on sets of tests that are properly validated, but wherein the proper validation of such tests can be difficult to obtain without making such tests and benchmarks more publicly available, thus increasing the likelihood that the very same test materials become available as the materials used to train the very same AI models they are intended to validate. Avoiding publicly available test descriptions reduces the likelihood that the tests are properly validated.

4.2. Frontier Benchmarks, Emergent Abilities, and Ongoing Challenges

Recent benchmarks like *FrontierMath* promise to avoid data contamination by creating new, unseen problems [31]. Although such efforts can provide more reliable assessments, these are often domain-specific (e.g., math problems) and may not capture other dimensions of intelligence, such as creativity or social reasoning. Moreover, while such tests may be appropriate and meaningful to use as tests of domain-specific knowledge, they can hardly be used to generalize to support any claims of AGI unless a link is made between these tests and an operational definition and measure of AGI or an extremely narrow definition of AGI is employed.

Moreover, tasks that appear novel to humans might still be partly predictable to a large AI model if they are structurally similar to data encountered during training. Evaluations must constantly evolve to stay ahead of model training corpora, implying a need for perpetual "benchmark refresh" to avoid the saturating phenomenon where top models max out old tests without truly scaling their general reasoning skills.

4.3. Moving Toward Adaptive, Psychometrically Informed Tests

Drawing on psychometric principles, a future AGI evaluation might include:

- **Adaptive Testing:** Dynamically adjusting question difficulty based on performance, akin to how human IQ tests find an individual's ceiling.
- **Construct Validity:** Designing tasks to isolate cognitive skills like abstraction, causal reasoning, and ethical judgment rather than conflating these with domain knowledge.
- **Standardized Administration:** Ensuring no training-data overlap and consistent testing conditions to rule out data leaks or resource inequalities.
- **Reliability Checks:** Re-testing models under different initialization or prompt conditions to gauge performance stability.

If coupled with open-source benchmarks and transparent reporting, such rigorous approaches could mitigate illusions of progress. They would also enable meaningful comparisons across diverse AI architectures and help identify truly generalizable skills.

5. Psychometric Perspectives

5.1. Validity, Reliability, and Human Comparison

Psychometrics emphasizes using validated constructs, consistent scoring, and normed data to measure latent traits such as intelligence or personality [32]. In AGI, these concepts remain underused.

 doi:10.20944/preprints202503.1581.v1

Few AI tests have the reliability or normative basis that let us confidently say, "System A has an overall intelligence factor of X."

One promising idea is to apply tests to humans and machines, calibrating the difficulty level to distinguish relevant cognitive abilities. For instance, if a model is tested on an exhaustive battery of tasks—linguistic, logical, social, creative—and its performance distribution is systematically compared to human data, we gain richer insight than a single pass/fail metric like the Turing Test. Achieving robust psychometric norms, however, requires extensive data collection and iterative refinement.

5.2. Sensitivity and Specificity in AI Tests

Borrowing from clinical diagnostics, psychometric tests for AGI should maximize *sensitivity* (correctly identifying accurate intelligence where it exists) and *specificity* (minimizing false positives). A highly sensitive test ensures that emergent but real capabilities are recognized, while a particular test reduces the chance that clever optimization strategies are misread as general intelligence [35,36].

When a model excels on a test, the question should be: does it generalize to variations of that test or related tasks? If minor tweaks cause performance to plummet, the system likely used brittle, domain-specific heuristics. This principle aligns with psychometric reliability, where consistent results across iterations or slightly perturbed scenarios indicate more genuine ability.

6. Ethical Implications

6.1. Risks of Overclaiming and Societal Misdirection

Exaggerated AGI claims can erode public trust and siphon resources from more urgent AI challenges, such as fairness or explainability [14,33]. Overhyping can also fuel unrealistic fears (e.g., near-term "superintelligence" takeovers) that prompt reactionary or misaligned policy responses. Public misconceptions about AGI, often fueled by sensationalist media or inflated marketing messages, risk creating a distorted understanding of AI's current and future capabilities. This can lead to a "mismatch of priorities," where societal focus shifts towards hypothetical, low-probability risks rather than addressing immediate concerns like algorithmic biases, unethical surveillance practices, or the environmental costs of training large-scale models.

In governance contexts, harmful decisions could result if a purportedly "AGI-level" system is deployed in critical domains (e.g., healthcare, defense) without genuine general intelligence. The consequences of such deployments could range from operational failures to catastrophic outcomes, particularly in high-stakes scenarios where human lives or global stability are at risk.

To prevent these outcomes, Ethical AI research must carefully bridge the gap between public perceptions of AGI—often shaped by marketing or media sensationalism—and the actual capabilities of deployed systems. This requires a multifaceted strategy, including increasing transparency in AI development, fostering critical public discourse, and engaging policymakers in understanding the nuanced limitations and risks of current AI technology.

Ultimately, inflated claims about AGI may not only hinder progress towards solving pressing AI challenges but could also undermine trust in the broader AI field, leading to a chilling effect on innovation and collaboration.

6.2. Transparency, Accountability, and Benchmark Honesty

To mitigate these risks, one approach is for AI developers to communicate their models' limitations and training scope. Openness about test contamination, domain coverage, and known failure modes is key.

Further, an independent review process—similar to an institutional review board—could evaluate high-stakes AI claims and ensure that benchmarks meet scientific standards [34].

6.3. On the Horizon

Should genuine AGI or near-AGI systems emerge, their societal impact will dwarf that of narrow AI. By forging robust benchmarks and psychometric methodologies now, the framework provided

 doi:10.20944/preprints202503.1581.v1

here lays the groundwork for detecting, validating, and safely integrating more powerful AI in the future. The ethical impetus is to ensure that gains in intelligence-like performance align with values such as well-being, justice, and transparency.

7. Conclusion

7.1. Summary of Key Findings

In this paper I have examined AGI from multiple disciplinary perspectives:

- **Conceptual Foundations:** Intelligence may be less a single capacity than a mosaic of specialized abilities, complicating efforts to replicate or measure "general" intelligence in machines.
- **LLM Introspection and Emergent Behaviors:** While large-scale models display intriguing phenomena—such as apparent self-reflection or sudden skill acquisition—these might be measurement artifacts or sophisticated pattern usage rather than genuine general intelligence.
- **Philosophical Critiques:** Searle's Chinese Room highlights the gap between simulating understanding and possessing it. The test-passing illusion warns against conflating performance metrics with true cognition.
- **Benchmark Limitations:** Existing AGI benchmarks often lack psychometric rigor and may be contaminated by training data, overstating model competencies.
- **Ethical Dimensions:** Overclaims about AGI risk misallocation of resources, erosion of trust, and misguided policy. Transparent, accountable practices are essential for responsible AI advancement.

7.2. Toward a More Rigorous AGI Paradigm

A consistent theme in this paper is the need for *conceptual clarity* and *methodological rigor*. Cognitive neuroscience and neuropsychology suggests complex, semi-modular, and network based intelligence rather than a holistic intelligence monolith. Large language models, while impressive, might only approximate portions of cognition through scale and optimization. Psychometrics provides valuable lessons for constructing and validating tests that measure underlying capabilities, not just surface performance. Philosophical debates remind us that function and understanding are not necessarily interchangeable. Finally, ethical guardrails ensure that the path to AGI, if it indeed exists, proceeds without undue harm or deception.

7.3. Future Directions

Moving forward, AGI research should:

- **Adopt Psychometric Standards:** Develop adaptive, validated, contamination-resistant benchmarks with strong construct validity and reliability measures.
- **Investigate Mechanisms in LLMs:** Probe how introspection-like and emergent behaviors arise, distinguishing illusions from substantive progress toward general cognition.
- **Incorporate Interdisciplinary Insights:** Use philosophy, cognitive science, and neuroscience to refine definitions of intelligence and create ethically robust AI.
- **Ensure Ethical Transparency:** Avoid overstating capabilities, reveal evaluation details, and engage with regulatory bodies when deploying high-stakes systems.

By pursuing these steps, the AI community can transition from rhetorical claims of "human-level" or "general" AI to empirically grounded assessments. The long-term vision of machines that reason, learn, and adapt across all domains remains a profound challenge. Yet by embracing honest benchmarking, interdisciplinary collaboration, and ethical accountability, we stand to maximize the benefits and minimize the risks on the winding road to genuine AGI—should it one day emerge.

References

1. Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial General Intelligence*. Springer.

 doi:10.20944/preprints202503.1581.v1

2. Legg, S., & Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, 17(4), 391–444.
3. Silver, D., Schrittwieser, J., Simonyan, K., et al. (2018). A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go Through Self-Play. *Science*, 362(6419), 1140–1144.
4.
5. Marcus, G. (2022). Artificial Intelligence: A Guide to the Future. [Publisher Information].
6. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Press.
7. Hernandez-Orallo, J. (2017). *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*. Cambridge University Press.
8. Gottfredson, L. S. (1997). Mainstream Science on Intelligence: An Editorial With 52 Signatories, History, and Bibliography. *Intelligence*, 24(1), 13–23.
9. Sternberg, R. J. (2000). The Modularity of Mind: Debates in Cognitive Science. In *Handbook of Intelligence* (pp. 15–37). Cambridge University Press.
10. Marcus, G. (2022). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *AI Magazine*, 43(1), 49–63.
11. Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3(3), 417–457.
12. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
13. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Dafoe, A. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228*.
14. Whittlestone, J., Nyrup, R., Alexandrova, A., & Dihal, K. (2019). Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research. *Minds and Machines*, 29(3), 395–410.
15. Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. *The American Journal of Psychology*, 15(2), 201–293.
16. Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books.
17. Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4), 245–313.
18. Gazzaniga, M. S. (2018). *Cognitive Neuroscience: The Biology of the Mind*. W.W. Norton & Company.
19. Silver, D., et al. (2017). Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *Nature*, 550, 354–359.
20. Brown, T., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
21. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building Machines that Learn and Think Like People. *Behavioral and Brain Sciences*, 40, e253.
22. Burkart, J. M., Schubiger, M. N., & van Schaik, C. P. (2017). The evolution of general intelligence. *Behavioral and Brain Sciences*, 40, e195.
23. [Placeholder: Example research on LLM introspection or self-reflection; replace with a real citation as needed.]
24. [Placeholder: Example research on privileged access in LLMs; replace with a real citation as needed.]
25. [Placeholder: Example reference discussing emergent abilities, e.g., DeepChecks blog or relevant arXiv.]
26. [Placeholder: E.g., a Wired article or academic paper calling emergent claims a “mirage.”]
27. Nass, C., & Moon, Y. (1994). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 50(1), 81–103.
28. Tihanyi, N., Bisztray, T., Dubniczky, R. A., Toth, R., Borsos, B., Cherif, B., Ferrag, M. A., Muzsai, L., Jain, R., Marinelli, R., Cordeiro, L. C., Debbah, M., Mavroeidis, V., & Josang, A. (2024). Dynamic Intelligence Assessment: Benchmarking LLMs on the Road to AGI with a Focus on Model Confidence. *arXiv preprint arXiv:2410.15490*.
29. WJARR Editorial. (2024). A Critical Review Towards Artificial General Intelligence: Challenges, Ethical Considerations, and the Path Forward. *World Journal of Advanced Research and Reviews*, 8(1), 1–12.
30. ResearchGate Authors. (2024). Opinion Paper: The Frustrating Quest to Define AGI. *ResearchGate Publications*.
31. [Placeholder: Reference for FrontierMath or other advanced new benchmark to avoid contamination.]
32. Cronbach, L. J., & Meehl, P. E. (1971). Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4), 281–302.
33. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

 doi:10.20944/preprints202503.1581.v1

34. Floridi, L., & Cowls, J. (2020). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 2(1).
35. Altman, D. G., & Bland, J. M. (1994). Diagnostic Tests 1: Sensitivity and Specificity. *British Medical Journal*, 308(6943), 1552.
36. Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.
37. Price, C. J., & Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends in Cognitive Sciences*, 6(10), 416-421.
38. Fodor, J. A. (2008). The modularity of mind: An essay on faculty psychology. In J. E. Adler & L. J. Rips (Eds.), *Reasoning: Studies of human inference and its foundations* (pp. 878–914). Cambridge University Press.
39. Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433–460.
40. Weizenbaum, J. (1966). ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1), 36–45.
41. Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), 277–301.
42. Riek, L. D. (2017). Healthcare robotics. *Communications of the ACM*, 60(11), 68–78.
43. de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2017). Trust in automation: An updated review of trust across disciplines. *Human Factors*, 59(5), 834–861.
44. Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219–232.