The Epistemic Limits of Artificial General Intelligence: Between Accumulated Knowledge, Scientific Discovery, and Informational Vulnerability

Zhivko Georgiev
G Consulting

Lyubomir Ivanov
Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

Abstract

This article argues that artificial general intelligence (AGI), even in a highly advanced form, should not be understood as an autonomous producer of truth independent of the quality, accessibility, and verifiability of available knowledge. A key distinction is made between AGI as a theoretical horizon and contemporary advanced AI systems, including large language models and tool-augmented or agent-based systems, which constitute the empirical basis of this analysis. The central claim is that, in the natural sciences—particularly at the frontiers of contemporary physics, chemistry, and biology—AGI may accelerate analysis, support hypothesis formation, and participate in scientific workflows, yet it cannot independently resolve competing theoretical frameworks in the absence of new empirical evidence and external experimental validation. In the humanities and social sciences, its role is likewise better understood as instrumental rather than sovereignly interpretive. The article further examines two major classes of risk: (1) AGI's epistemic dependence on a reliable flow of information; and (2) the possibility that these same technologies may amplify the capabilities of malicious actors, particularly in cyberspace. These risks are not external to the main thesis but follow directly from it.

Keywords

artificial general intelligence; epistemology; large language models; scientific discovery; social sciences; disinformation; cybersecurity; quantum technologies

1. Introduction

Public debate on artificial intelligence is often structured between two extremes: technological maximalism and radical skepticism. According to the former, AGI is on the verge of replacing the human scientist, expert, and analyst. According to the latter, it will inevitably remain no more than a mechanism for the statistical rearrangement of already existing knowledge. A more plausible position lies between these poles. Contemporary AI systems are already transforming scientific and professional labour, not as independent intellectual agents, but as infrastructures for accelerated analysis, synthesis, and search across large bodies of information (Messeri and Crockett, 2024; Zhang et al., 2025). Therefore, in the following discussion, the term AI is used cautiously: not as a claim that such a fully general and autonomous system already exists, but as a horizon against which the possibilities and limitations of today's advanced AI systems can be explored. Central to understanding those possibilities and limits is the distinction between processing knowledge and producing validated new knowledge.

2. Theoretical Framework: AGI and the Problem of Truth

The central thesis of this article is that AGI should not be conceived as an autonomous arbiter of truth. Its cognitive power remains structurally dependent on pre-existing data, models, and procedures of external verification. If knowledge is understood not merely as the accumulation of facts, but as content that has been tested, contextualized, and interpreted, then access to vast information cannot be equated with access to truth (Kalai et al., 2025; Messeri and Crockett, 2024). Contemporary scholarship on AI and the scientific method underscores precisely this point: large language models may support hypothesis generation, analysis, and discovery-oriented work, but their contribution remains reliable only when embedded within a broader cycle of human purposes, measurement, experimentation, and evaluation (Zhang et al., 2025). The argument developed here is therefore aimed less at a metaphysical denial of machine intelligence than at a more limited epistemic claim: without sustained links to observation, experiment, and institutional validation, even the most

capable AI remains dependent on knowledge whose truth it has not independently established.

3. Competing Hypotheses and the Limits of Textual Intelligence

Quantum mechanics provides an instructive example. Despite its predictive success, its interpretations remain contested. A 2025 *Nature* survey indicates that physicists continue to be sharply divided over what quantum mechanics actually says about reality (Gibney, 2025). This highlights the limits of purely textual and logical analysis. In such contexts, AI systems may organize arguments and identify gaps, but cannot independently determine which interpretation is correct without new empirical evidence.

A fundamental limitation therefore comes into view. When scientific controversy concerns competing hypotheses that remain partially compatible with existing data, what proves decisive is not further textual recombination, but new facts, new measurements, and new experimental regimes. A system whose primary epistemic interface is symbolic/digital rather than sensorimotor would not, in isolation, generate such facts. Accordingly, insofar as AGI operates primarily on already available digital knowledge, it is more accurately described as an accelerator of research than an epistemic arbiter of competing hypotheses.

4. Can AGI Nevertheless Participate in Discovery?

The stronger version of the skeptical thesis—that artificial intelligence cannot contribute to scientific discovery—is no longer fully defensible. There is now substantial evidence that AI systems can make meaningful contributions to the scientific process. AlphaFold, for example, has achieved highly accurate predictions of protein structures (Jumper et al., 2021), while the Co-Scientist system demonstrates that a language model equipped with search tools, code execution, documentation access, and laboratory automation can design and perform complex experimental tasks (Boiko, MacKnight and Gomes, 2023).

This does not mean that AI replaces the scientist. Rather, it suggests that when AI systems are connected to instruments of observation and verification, they can become meaningful participants in the cycle linking hypothesis, experiment, and result. The more precise formulation, therefore, is not that AGI is permanently confined to the role of archivist of past knowledge, but that its ability to move beyond existing knowledge depends on integration with experimental systems and instruments (Zhang et al., 2025; Boiko, MacKnight and Gomes, 2023). Without this, AI remains a system of intelligent recombination; with such integration, it may contribute to the discovery of new regularities.

5. Natural Sciences versus the Humanities and Social Sciences

In the natural sciences, the performance of AI is often easier to assess because these domains more frequently involve clearly defined objects, standardized data, and robust procedures of verification. In the humanities and social sciences, by contrast, the situation is considerably more complex, since the subject matter includes meanings, values, historical context, cultural specificity, and linguistic ambiguity. The difference is one of degree rather than kind. Natural sciences offer clearer validation procedures, while social sciences involve greater contextual complexity. Both domains include elements of uncertainty and empirical grounding which, however, does not render AGI useless in such fields. Research on the use of large language models in law, for instance, points to substantial applications in document drafting, legal research, case analysis, compliance, and education (Dehghani et al., 2025). Yet those same studies also emphasize the need for accountability, verification, and human oversight.

A similar pattern emerges in social simulations. Research indicates that large language models may be used for role-based modelling and computational social experiments, but it also reveals inconsistencies in simulated roles and a weak correlation between a model's general "power" and the reliability of its behaviour in social contexts (Huang et al., 2024). In sociology, political science, psychology, and

macroeconomics, AGI may therefore prove useful for processing large datasets and for scenario modelling, but it should not be regarded as the ultimate bearer of contextually valid understanding.

6. Data, Context, and the Problem of Reliability

The principal resource of AI systems is information, but not all information is reliable knowledge. Contemporary research warns that AI in science may generate what has been described as “illusions of understanding”, namely situations in which the quantity of output increases without a corresponding increase in genuine explanatory insight (Messeri and Crockett, 2024). This is an important epistemic problem, because a model may appear highly persuasive without actually being trustworthy.

To this must be added the risk of so-called *model collapse*. A 2024 *Nature* publication shows that when generative models are recursively trained on data produced by earlier generations of models, degradation and a resulting “misperception of reality” may occur (Shumailov et al., 2024). More broadly, this suggests that an informational environment saturated with synthetic and unreliable content may do more than merely hinder AI systems; it may gradually erode their cognitive reliability. Concerns that AGI may be vulnerable to systematic disinformation and to what might metaphorically be described as “cognitive viruses” therefore have a clear analogue in the literature on data poisoning, synthetic contamination of training corpora, and recursive model degradation (Messeri and Crockett, 2024; Shumailov et al., 2024).

An important practical conclusion follows. If AGI’s access to reliable data is restricted while the public sphere becomes saturated with plausible but difficult-to-verify disinformation, then its cognitive power will depend increasingly on the quality of its built-in filters and on procedures of external validation. For that reason, the claim that strategically important scientific, governmental, and corporate information is unlikely to remain fully open once sufficiently powerful AI systems become widespread is institutionally plausible, even if the precise scope and mechanisms of such restriction cannot yet be

determined with certainty. It is this information vulnerability that provides the transition to the second major concern of this work: if epistemic efficiency depends on the quality of information, then the same digital environment that supports advanced AI can also become a field of increased strategic manipulation, including in the field of cybersecurity.

7. AGI, Cyber Threats, and the Asymmetry of Resources

The second major class of risks concerns not whether AGI will somehow "escape control" in an abstract sense, but who will use such systems and to what ends. Cybersecurity scholarship already examines the ways in which large language models may lower the threshold for the planning, automation, and execution of attacks, including through reconnaissance, coding assistance, and the scaling of malicious operations (Ayzenshteyn, Weiss and Mirsky, 2024). It is therefore entirely justified to worry that the benefits of AGI will not be distributed symmetrically. High-resource actors will be better positioned to invest in defence, while weaker states, smaller organizations, and individual users may become disproportionately vulnerable.

This risk is sharpened further by the quantum factor. NIST is already officially working on the transition from vulnerable cryptographic standards to quantum-resistant ones, indicating that the threat posed to parts of today's cryptographic infrastructure is not merely speculative, but institutionally recognized (Moody, 2024). If increasingly capable AI systems are added to this environment and used to automate aspects of cyberattack, then the combination of AI and quantum pressure on security may create a qualitatively new level of risk, particularly for actors lacking sufficient technological and financial capacity for defense.

8. Counterargument: Is Autonomous Self-Improvement Possible?

A serious counterargument to the skeptical position holds that future AI agents may transcend current limitations through self-improvement, inter-agent coordination, and ever deeper integration

between cyberspace and the physical world, with AGI's epistemic ceiling partly determined by its degree of coupling to physical reality, and therefore varying across different system architectures and levels of embodiment. This argument should not be dismissed prematurely. Contemporary developments in agentic AI, tool-augmented models, and semi-autonomous scientific systems clearly show that the boundary between a model that merely responds and a system that acts is gradually shifting (Zhang et al., 2025; Boiko, MacKnight and Gomes, 2023).

At the same time, while systems are becoming more complex, there is no evidence supporting the conclusion that these developments have already produced an independent, fully uncontrolled, and epistemically autonomous AGI. What is emerging, rather, is movement toward increasingly complex sociotechnical assemblages in which human control, infrastructural constraints, and institutional regulation remain indispensable (Zhang et al., 2025; Messeri and Crockett, 2024).

9. Conclusion

The principal conclusion of this article is that AGI should be understood neither as a magical substitute for human intelligence nor as a useless statistical automaton. Its real power is considerable, but structurally constrained. In the natural sciences, it can significantly accelerate research, yet without new data and external validation it cannot independently resolve competing hypotheses. In the humanities and social sciences, it is valuable as a tool for processing, classification, support, and modelling, but remains limited wherever context, meaning, and historical depth are decisive. These conclusions apply most certainly to today's advanced AI systems and, by analogy, to the plausible near-term trajectories often associated with AI; they should not be interpreted as a definitive refutation of any more powerful future form of machine intelligence.

The most significant risks therefore lie not necessarily in a hypothetical "machine rebellion", but in AGI's dependence on reliable information, its vulnerability to disinformation, the emergence of recursive

feedback loops between AI-generated and human-generated information, and the possibility that these same technologies may increase the productivity of malicious actors. Furthermore, historical development in such contexts is often non-linear, with relatively gradual improvements giving rise to sudden shifts in capability, institutional structure, and the distribution of epistemic authority. The future of AGI is, for that reason, not only a technological question, but also an epistemic, institutional, and geopolitical one.

References

Ayzenshteyn, D., Weiss, R. and Mirsky, Y. (2024). The Best Defense is a Good Offense: Countering LLM-Powered Cyberattacks. *arXiv*:2410.15396.

Boiko, D.A., MacKnight, R. and Gomes, G. (2023). Autonomous chemical research with large language models. *Nature,* **624**, 570–578.

Dehghani, F., Dehghani, R., Ardebili, Y.N. and Rahnamayan, S. (2025). Large Language Models in Legal Systems: A Survey. *Humanities and Social Sciences Communications*, **12**, Article No 1977.

Gibney, E. (2025). Physicists disagree wildly on what quantum mechanics says about reality, *Nature* survey shows. *Nature,* **643**, 1175–1179.

Huang, Y. et al. (2024). Social Science Meets LLMs: How Reliable Are Large Language Models in Social Simulations? *arXiv*:2410.23426.

Jumper, J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature,* **596**, 583–589.

Kalai, A.T., Nachum, O., Vempala, S.S. and Zhang, E. (2025). Why Language Models Hallucinate. *arXiv*:2509.04664

Messeri, L. and Crockett, M.J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature,* **627**, 49–58.

Moody, D. et al. (2024). *IR 8547:* Transition to Post-Quantum Cryptography Standards. National Institute of Standards and Technology (NIST).

Shumailov, I. et al. (2024). AI models collapse when trained on recursively generated data. *Nature,* **631**, 755–759.

Zhang, Y. et al. (2025). Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence,* 1:14.