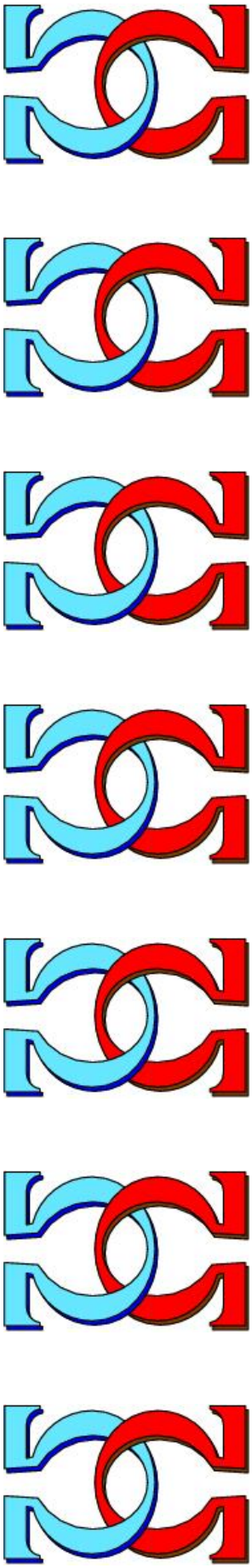

CDMTCS
Research
Report
Series

A Critical Analysis of the Status and Impact of Artificial Intelligence

Joseph Sifakis

Grenoble-Alpes University, Verimag Laboratory

CDMTCS-589
May 2026

Centre for Discrete Mathematics and
Theoretical Computer Science

A Critical Analysis of the Status and Impact of Artificial Intelligence

Joseph Sifakis,
Grenoble-Alpes University, Verimag Laboratory

Abstract

Artificial intelligence is not just a technical discipline—it is a force that actively reshapes the cognitive and social world it inhabits. This unique characteristic creates profound epistemological challenges for validating AI systems and explains much of the public confusion surrounding their purpose and limits.

Despite remarkable advances, AI remains a discipline in its infancy. Its most significant applications—particularly monitors and autonomous systems—lie ahead, blocked by foundational gaps that current approaches cannot bridge. Scaling parameters and chasing Artificial General Intelligence are distractions from the hard problems that must be solved.

The development of AI is shaped by forces beyond science: large technology companies, academic institutions, government policies, and public intellectuals all compete to steer its direction. Their interplay, not technical progress alone, will determine AI's societal role.

This article calls for clarity over hype, scrutiny of concentrated power, and collective responsibility. The future of AI is not fate—it is a choice. And the choice is ours.

1. A Discipline Apart: The Technical, Cognitive, and Social Dimensions of AI

AI stands apart from other disciplines. Natural sciences examine an objective reality; social sciences analyze human behavior without fundamentally altering it. AI constructs systems that actively remake the cognitive and social world they inhabit.

Consequently, the debate about AI cannot be confined to a strictly technical or scientific framework. It inevitably involves subjective factors and external forces: public opinion, industrial and financial interests, state institutions, the media, and public intellectuals all compete to steer its development.

This unique position creates a conundrum. On one hand, AI is an intensely complex technical field. On the other, it presents itself to the world through a deceptively familiar lens. This gap between appearance and reality profoundly influences public opinion and attitudes toward AI and creates considerable confusion in the public debate.

The discussion about the nature of AI, its purpose, and its direction is inevitably influenced by the economic and social issues it raises. How does it compare to human intelligence? What avenues should be explored and what types of applications should be focused on? How can the trustworthiness of AI systems be guaranteed? The answers to these questions will determine the limits in the mechanistic production and application of knowledge, but their value will depend on our ability to use technical results to build applications that are recognized as useful and accepted by society and its institutions.

Intelligence is a multifaceted concept. It is meaningless to say that "system X is more intelligent than system Y" unless we specify the type of task they are designed to perform and the criteria for success. The usefulness of AI will be judged on its ability to mimic cognitive functions and, ultimately, on its ability to *replace* humans in performing a wide variety of tasks. Human intelligence is not a theoretical abstraction. It is the product of a long evolution in interaction with a given physical

environment; it would have taken other forms if we had emerged on a planet different from our own. So, if AI leads us to build agents that can perform extraordinary feats in a video game set in a world different from our own, this has value as a scientific achievement but limited relevance to the world we actually inhabit.

Validating the results of an AI system poses profound epistemological problems. When a system plays chess, the method is simple: we judge the results against objective rules. But when a system converses, makes recommendations, or takes decisions, validation becomes unclear. Should we use social science methods to evaluate the user experience or to measure societal impact? The answer is that AI is a hybrid discipline that requires a hybrid epistemology. Nevertheless, this very fact creates a vulnerability. When validation modes are multiple and contested, it becomes possible to choose the mode that suits one's own discourse. Validating AI requires not only technical knowledge, but also a clear vision of how we know what we claim to know.

In this essay, we examine AI not only in its scientific and technical dimension, but also from the perspective of its societal impact, taking into account the extra-scientific forces that shape it. We argue that AI still has a long way to go before it becomes a mature discipline, and that the most significant applications—particularly autonomous systems—remain ahead of us. We advocate the need to dispel the mystification, understand the real forces at play, and emphasize our collective responsibility in managing this technology for the good of society.

2. The Low Epistemic Barrier: Confusion, Myth, and Misinformation

Public opinion on the role of AI is currently mired in confusion and overwhelmed by misinformation. This is facilitated by the very nature of the discipline and perpetuated by actors who have every interest in preventing a rigorous and informed public debate.

2.1 How Confusion Takes Hold

AI speaks to the most universal of human attributes: intelligence, knowledge, creativity, and decision-making. It is therefore inevitable that it attracts interest—and opinion—from all quarters. Compounding this is our innate tendency to project human qualities onto things that appear to act with intent. When an AI writes a poem or holds a conversation, it is easy to conclude that it "thinks", "feels" or "understands". This relieves us of the need to grasp the complex statistical mechanisms of AI systems, replacing them with a comfortable, human-centric narrative.

The field itself encourages this confusion. AI borrows heavily from the vocabulary of human cognition: neural networks, learning, memory, attention, understanding. These are powerful metaphors, but they do not literally describe what the technology does. A neural network is not a brain; it is a complex mathematical function. The word "understanding" in AI is a conceptual minefield. It is a linguistic shortcut that allows elaborate arguments to be constructed on metaphorical foundations without ever touching technical ground. For the ordinary user, moreover, the technology is designed to be invisible. It presents a fluid, simple interface that masks an infrastructure of breath-taking complexity. Ease of use creates an illusion of ease of understanding.

This illusion is compounded by a deeper structural feature: unlike mathematics, physics, or traditional engineering disciplines, AI has a remarkably low "epistemic barrier." It is very easy to talk with an air of false expertise about superintelligence and the risks it poses without having a deep understanding of the nature of this technology. This creates a perverse incentive: the more easily understandable—and often misleading—the discourse, the more readily it is accepted and amplified by the media.

2.2 Myths and Reality

The debate on AI attracts now far more public interest than debates on the Big Bang or the Origin of Species—and for good reason. Like any other major debate, it is caught between two opposite poles: on one side, the presentation of technical details so dense they repel the audience; on the other, an excessive simplification that is easy to grasp and pleasant to hear.

Unfortunately, it is the second approach that has overwhelmingly prevailed. Faced with a technologically complex reality whose implications radiate in every direction, no public debate worthy of the name has emerged. AI is perceived by the general public through a mystifying and irrational discourse, one in which facts and technical challenges are overshadowed by the raw power of myths. Such mystification prevents the public from forming a realistic opinion about the very nature of AI and from taking a rational approach to the issues raised by its use.

2.3 The Problem of Authority

AI's multiple facets and inherent accessibility make it unusually difficult to distinguish authoritative voices from charlatans. The result is a disordered public debate inundated with flattering rhetoric that masks real problems behind uncritical optimism.

Charlatanism takes many forms: researchers chasing publication counts, consultants posing as experts, spokespersons for technology companies. Some produce wish lists of problems followed by empty assurances that solutions will come. Others operate at the "visionary" level, showing little interest in the engineering rigor that real-world systems demand. A proliferation of some "AI ethicists" offer brilliant ideas but lack the technical expertise to conduct ethical analysis.

Demagoguery often accompanies charlatanism, advocating simplicity over rigor. The message is seductive: you don't need technical understanding—just a clever trick, a perceptive mind. "Ingenious" experiments are overinterpreted as breakthroughs, making everyone feel like an expert while genuine understanding recedes.

The incentive structure is blatantly clear: sensational narratives about AI sentience, deception, or imminent AGI attract clicks and funding. Technically grounded discussions of limitations are ignored. Charlatans thrive because they tell audiences what they want to hear, and the system rewards them for it.

3. Quo Vadis AI? A Colossus with Feet of Clay

AI today is a colossus with feet of clay. Behind these remarkable achievements lie gaps in the theoretical foundations that would enable a better response to the technical requirements necessary for their application.

We examine those gaps and argue that the field's most significant applications— monitors and autonomous systems— are still far from being realized, pending the progress needed to fill these gaps.

3.1 Beyond Conversation: Monitors and Autonomous Systems

The current situation requires clarification of the very nature of AI and the direction it should actually take. AI currently focuses on one type of system: conversational agents that interact with users in a question-and-answer mode such as, chatbots, writing assistants, and image generators. But an analysis of genuine societal needs reveals two other categories of intelligent systems, each with fundamentally different characteristics and requirements.

Monitors. Unlike conversational systems, monitors do not engage in dialogue. They operate as continuous observers of streaming data, detecting patterns, anomalies, and trends in real time with minimal latency and high reliability. Their role is not conversation but insight: predictions, alerts, diagnoses. They can anticipate traffic fluctuations in a metropolitan area, forecast energy consumption in electrical grids, or locate failures in industrial systems through root-cause analysis. We need reliable smart monitors to effectively control and manage resources and services.

Autonomous systems. Beyond monitors, we need autonomous systems capable of delivering guaranteed-reliability services without human intervention. These are composed of multiple agents, each pursuing its own objectives, but must coordinate to satisfy overall system goals. Unlike conversational systems, autonomous systems must pursue goals over time in dynamic, partially observable environments. The potential applications are numerous: autonomous transportation, smart grids, smart factories, self-managing communication networks. Realizing this vision poses technical challenges that go well beyond the current state of the art [1].

The history of autonomous vehicles is a cautionary tale. Despite confident predictions from industry leaders—"fully autonomous vehicles by 2020", "1 million robotaxis by 2020"—the technology remains stubbornly distant [2]. What we have learned is that navigating the open world requires forms of reasoning and reliability that statistical pattern matching, no matter how sophisticated, cannot provide. Similarly, the bold declarations that "2025 will be the year of AI agents" are already proving hollow [3].

We need AI systems that are not only capable of learning, but possess genuine reasoning abilities—systems that can pursue goals rationally and in accordance with technical, legal, and ethical standards. Above all, they must be able to perform critical tasks while ensuring levels of reliability that current AI cannot achieve: 10^{-8} failures per hour of operation [4]. To address these shortcomings, we must go beyond simply adjusting parameters and instead create fundamentally new architectures that integrate learning with reasoning, situational awareness and planning.

Building autonomous systems capable of reliably replacing human agents in complex organizations is the ultimate goal of AI. This is a concrete, experimentally validatable goal—one with clear metrics, known challenges, and genuine societal value. As explained in the following section, this is very different from AGI.

3.2 The Foundational Gap: Missing Building Blocks

3.2.1 Important Clarifications

The scientific community, probably dazzled by rapid progress, has neglected a foundational task: establishing a rigorous framework that distinguishes machine learning systems from the traditional computational artifacts with which they are constantly compared. This negligence has produced a confusion so pervasive that it infects even the most advanced research.

Two Computational Paradigms

Two radically different computational paradigms exist, mirroring Kahneman's distinction between fast and slow thinking [5]. Neural networks, like System 1, operate quickly, associatively, and opaquely—they deliver answers without revealing the process by which those answers were reached. Through algorithmic learning, they adjust parameters to interpolate appropriate responses in a multidimensional space. On the other hand, traditional software and hardware, like System 2, operate procedurally, in a manner that can in principle be traced and verified using well-established techniques.

The Black Box Question

Neural networks are often called "black boxes", but this term can be understood in two ways. To simplify matters, let's consider a non-recursive neural network. It can be represented by a deterministic function: for a given input vector, it produces a corresponding output vector. This function could theoretically be calculated by composing the functions of individual neurons. But such a formal calculation is impossible, not because of a theoretical limitation, but because of explosive complexity. It is this complexity that implies the non-explainability of neural networks. It prevents us from obtaining precise behavioral models, as we can do by analyzing software or hardware systems.

Unfortunately, the uncritical transposition of concepts from traditional systems engineering is an additional source of confusion. Calling neural networks “models” introduces misunderstandings that are found in a large number of works. Neural networks do not have the explanatory power of models used in scientific disciplines, which maintain a clear distinction between a physical system or an artifact and its representation. Software is not a model. To verify and analyze its behavior, we build models that are transition systems based on its operational semantics.
Furthermore, neural networks are often confused with the algorithms used to train them. Many people believe that because neural networks are produced by algorithms, they themselves can be modeled and analyzed mathematically. This is a category error. The *training algorithm* is a well-defined procedure. The *trained neural network* that results from this procedure is an artifact whose internal representations and decision boundaries emerge from the interaction of data, structure, and optimization dynamics. The general characterization of these dynamics remains an open question. To view the network as if it were simply executing its training algorithm is to confuse process with product.

The Interpretation Problem

When neural networks are applied to unstructured data—images, audio, natural language—a prior step is required that is never neutral: the data must be encoded into vectors. This encoding is itself an interpretation. It selects what counts as relevant, imposes a structure on the input, and inevitably discards information. The behavior of an image analysis system, therefore, is not merely the behavior of the neural network; it is the behavior of the network *plus* the interpretive choices embedded in the encoding. This interpretative pre-processing stage, which cannot be fully formalized, forecloses any possibility of rigorous analysis.

3.2.2 Technical vs. Non-Technical Systems: A Foundational Distinction

This discussion leads to a fundamental distinction rarely made in the literature: the distinction between *technical* and *non-technical systems*.

A technical system has inputs and outputs whose domains are formally defined, giving them unambiguous semantics. Its behavior can be rigorously characterized by relationships between possible input and output sequences. Its properties can be tested and, when possible, verified on behavioral models. A thermostat, a flight controller, or a chess-playing program are technical systems. Starting from an initial chessboard configuration, the chess game system generates responses based on the opponent's moves, a relationship that can, in principle, be formally characterized because we fully understand the rules of the game.
On the contrary, an LLM is a non-technical system. Its input/output relationship cannot be rigorously characterized because the domain (natural language) is semantically ambiguous. To determine that the correct answer to the question “What is the capital of France?” is “Paris”, you must first interpret the input/output correspondence before relating it to the reality of the world.

3.2.3 The Two Faces of Trustworthiness: Behavioral and Cognitive Properties

The trustworthiness of AI systems is a critical issue for their use particularly in systems that operate on data flows without human intervention. Despite the numerous studies devoted to this concept,

there is currently considerable confusion regarding its meaning for AI systems and the methods for guaranteeing its characteristic properties. The trustworthiness of AI systems encompasses two fundamentally different types of properties.

Behavioral Properties

Behavioral properties are properties whose validity depends solely on a system's input-output behavior. They include risk-related properties—safety (avoiding harmful states) and security (resilience against malicious actions)—and progress-related properties, such as performance, efficiency, and usability.

For technical AI systems, behavioral properties can only be validated through testing. Formal verification techniques are impossible to apply in the absence of faithful behavioral models. This limitation to testing significantly restricts our ability to accurately estimate and guarantee the trustworthiness of AI systems. It should be noted that for traditional critical systems, the strong reliability guarantees required by standards, are obtained through analysis methods that rely heavily on their behavioral and architectural models.

Today, we are sorely lacking rigorous testing methods that allow for a systematic exploration of the behavior of the system under test and the production of reproducible results.

A telling example is the validation of autonomous vehicles. At what point is an autonomous car safe enough? Given the stakes, the public debate rarely engages with the actual foundations of systems engineering. We are often presented with arguments based on miles driven—millions of miles in real traffic or billions in simulation—as proof of safety. Technically, this is not enough. Safety cannot be demonstrated by counting miles; it requires demonstrating that all possible dangerous situations have been adequately covered. This demands coverage criteria, as in traditional systems testing, yet such criteria are almost entirely absent for autonomous vehicles [6].
Another common argument compares autonomous driving systems favorably to human drivers, highlighting cases where the system outperforms the human. What this comparison omits are the countless situations handled effortlessly by the average driver but where autonomous driving systems fail catastrophically. The asymmetry is never acknowledged.

For AI systems, *safety* is the most important behavioral property. As such, it is the subject of numerous scientific publications and expert reports. It is also a key concern at international intergovernmental summits and institutions such as the UN [7]. Nonetheless, although the technical challenges to be overcome are often recognized, there remains a remarkable optimism that AI systems can achieve the same level of safety as traditional critical systems. It is particularly worrying that international experts do not recognize the obvious limitations and advocate adapting formal methods to neural systems, which is simply impossible in the absence of precise behavioral models.

Cognitive Properties

To define the trustworthiness of AI systems, we need a second type of properties that do not apply to traditional systems. Most AI systems are designed to mimic human behavior and must therefore satisfy human-centric, *cognitive properties*. These are essential for non-technical systems such as LLMs, which perform cognitive functions. They depend not only on observed input-output behavior, but also on the system's knowledge and how it applies that knowledge to make decisions.

The study of cognitive properties has given rise to a wealth of literature on “responsible AI”, “aligned AI”, and “ethical AI” [8]. In addition, these terms feature prominently in the marketing campaigns of major technology companies, which are supposed to strive to make them the essential attributes of their products.

Unfortunately, all this talk about cognitive properties seems to miss their true nature. Current studies evaluate cognitive properties based on behavioral criteria [9]. But assigning responsibility requires at least reasoning and intention capabilities, i.e., making deliberate decisions. Currently, AI systems have neither of these capabilities.

Responsibility, like all cognitive properties, requires that the system be aware of what is right or wrong, true or false. If a system claims that “the Earth is flat,” we have no way of determining whether it is lying or simply ignorant—because we have no way of determining what it actually “knows”.

A rigorous definition of cognitive properties will likely require architectural models whose behavior is governed by explicit knowledge stored in memory.

These observations raise a critical question about AI trustworthiness, particularly for AI agents. Many studies compare machine intelligence to human intelligence using various benchmarks—all of which are Turing-type behavioral tests. We must question the legitimacy of the conclusions drawn from these tests.

To conclude this discussion on cognitive properties, consider an autonomous driving system that passes the standard tests given to human drivers. Would this be sufficient to authorize it to drive? The obvious answer is no. But analyzing *why* this test is insufficient leads us to consider the cognitive properties we assume for humans but cannot guarantee for artificial agents.

A human driver who passes the driving test is assumed to possess a mental model of the world, an understanding of traffic rules, awareness of other drivers' intentions, and the capacity to make ethical judgments in unforeseen circumstances. An AI that passes the same behavioral test possesses none of these. It has simply demonstrated that its outputs can, under controlled conditions, mimic those of a competent driver. The difference between behavioral mimicry and true cognitive reliability is not a matter of philosophical abstraction. It is the difference between a system capable of functioning under specific conditions and a system that can be relied upon in any context—one that takes into account technical rules and also legal and ethical standards in the pursuit of its goals.

3.3 The Multi-Agent Gamble

Multi-agent systems will be the central focus of AI research in the coming years. There is a general consensus that learning alone is not enough to create autonomous agents. These agents must be equipped with the ability to reason based on a model of their "world". In this regard, a first step is bridging the gap between unstructured and structured data—particularly connecting natural language to semantically well-founded *world models*. A second step involves developing effective reasoning techniques to maintain consistency in updating world models, and establishing rational deliberative processes that enable goal planning and achievement. A third step, perhaps the most difficult, consists of developing methods for validating cognitive properties based on the analysis of a model that reflects how an agent manages knowledge and makes decisions.

Given these challenges, it is not surprising that attempts to create trustworthy AI agents have failed. In their frantic race to dominate the market, large technology companies quickly proposed protocols such as ACP, MCP, and A2A to create multi-agent systems and connect agents to tools [10]. But all this haste masks a fundamental reality: getting autonomous agents with mediocre reliability to cooperate increases the risk of failure exponentially.

Let’s take the example of RAG architectures, which are commonly used to build agents designed to process unstructured data using LLMs and in-memory data. The solutions currently being tested struggle to establish a reliable link between text queries produced by an LLM and structured queries (in SQL) to retrieve relevant data. Assuming an optimistic accuracy of 90% in the current state of the art, for text-to-SQL conversion, the success rate after only 10 successive queries drops to 34.8%.

3.4 Rethinking the Road: Challenges and Directions

We have presented a vision of AI that differs radically from the prevailing discourse, clearly highlighting some fundamental weaknesses in most existing work. Of course, the weaknesses pointed out do not detract from what has been achieved. But we must also recognize what they are not: they have failed so far to provide a solid foundation for the discipline. AI is still in its infancy. We only have the most basic elements and do not have the methods necessary to build trustworthy intelligent systems capable of rivaling human intelligence.

We have argued that *autonomy and trustworthiness* are the true ultimate goals—complementary and inseparable—especially for autonomous agents. These goals can only be achieved if we commit to analyzing them in depth and finding ways to overcome the inherent limitations of machine learning by integrating external knowledge sources and reasoning mechanisms. A clear-eyed analysis of the current state of the art shows that we are very, very far from achieving them.

4. Driving Blind: Big Tech's Strategy and Its Consequences

The research and technical teams at large technology companies, thanks to their critical mass of computing and human resources, as well as their talent, are the undisputed leaders, the most advanced and at the forefront of AI. This is changing the focus and priorities of the entire field, shifting it from scientific and public policy considerations to commercial and economic interests, with profound consequences discussed below.

4.1 Gigantism: A Profitable, Technically Flawed Strategy

Scaling laws have empirically shown that larger models improve performance, but these gains follow diminishing returns: computational cost grows exponentially while performance improves marginally. Some experts argue we are approaching the "end of the scaling era," constrained by data, energy, and hardware limitations [11].

Another argument has been put forward to justify continuing to scale up: sufficient growth in these parameters would enable reasoning capabilities comparable to those of humans. This argument is technically questionable—how can statistical simulation produce robust, generalizable reasoning? Leading researchers are divided, yet some players treat it as settled fact because it serves commercial interests.

The "bigger is better" hypothesis is not merely technical. It is a business strategy disguised as science, justifying billion-dollar investments under the guise of fundamental research. It also centralizes power: only those with enormous capital can compete, creating a structural "compute divide" in the field.

Most damagingly, this premise forecloses alternative research directions. By committing exclusively to scaling, the field has starved approaches—symbolic AI, neuro-symbolic systems, hybrid architectures—that might prove more efficient, interpretable, and reliable. The hypothesis has become a self-fulfilling prophecy: investment follows belief in scale, and the resulting concentration of power makes alternative paths increasingly difficult to imagine or pursue.

4.2 Appeal Over Utility -- The Consumerization of AI

A second pillar of the strategy of leading technology corporations is the systematic prioritization of *appeal over utility*. The public is offered not the most useful applications, but the most appealing and acceptable—tools designed for mass adoption rather than genuine social value: chatbots, image generators, and writing assistants.

The market is already validating this distinction. Meta has made this strategy explicit. As Chief Product Officer Chris Cox told employees, "We're going to go focus on entertainment, on connection

with friends, on how people live their lives" rather than competing on productivity [12] —a frank admission that keeping users engaged matters more than enabling them to accomplish tasks.

Meanwhile, the true transformative potential of AI—in scientific discovery (drug development, materials science, climate modeling), industrial automation (robotics, supply chain optimization), and the real economy—remains largely untapped. These applications are harder to sell. They require deep integration, specialized models, high reliability, and demonstrable return on investment. This is slow, unglamorous work. Commercial logic therefore dictates focusing first on the faster, broader consumer market, leaving the more impactful industrial and scientific applications to develop at a low pace.

Gartner predicts that over 40% of agentic AI projects will be canceled by the end of 2027 due to escalating costs, unclear business value, and inadequate risk controls, highlighting the gap between current hype and the technology's maturity [13].

Large technology companies have been quick to claim their share of the pie, offering protocols for multi-agent systems whose real-world applicability leaves much to be desired. But the gap between marketing narrative and technical reality remains as wide as ever.

All indicators suggest that, for the foreseeable future, the strategy of tech giants will focus primarily on general-purpose conversational AI: building customer loyalty through services that extend the web and replicate its business models.

4.3 AGI – The Swiss Army Knife of AI

Large technology companies excel at generating media hype commensurate with their economic weight, thereby serving their direct commercial interests while promoting a conquering technological vision. With a compliant cohort of experts drawn from academia, media, and think tanks, they promote the idea that AGI is the ultimate goal of AI—thereby narrowing the field's ambitions to a single, questionable, and fuzzy target.

AGI has proven to be the ultimate marketing tool, the most powerful slogan in the tech world. It captures the imagination of the public, investors, and top talent. It transforms a company from a simple product manufacturer into a player shaping the future of humanity. This narrative justifies everything: computing costs, data collection, ethical shortcuts. It is a license to act with messianic intent.

The discourse on AGI also stimulates investment and attracts talent. The brightest researchers want to work on the most difficult and important problems. The promise of working toward AGI is a recruitment tool that universities cannot match.

This phenomenon is all the more striking given that no one can say exactly what AGI is. The stated goal—to surpass humans in the performance of all (most?) intellectual tasks [14]—presupposes a fixed inventory of human capabilities. But no one knows how many different tasks characterize human intelligence, or whether such a list is even possible. So, AGI is described as a sort of Swiss Army knife of AI: a tool capable of doing anything in theory, but nothing without a user who can coordinate its functions in a targeted manner. Of course, machines may outperform humans in carrying out a set of tasks considered in isolation, but combining these tasks requires reasoning and coordination abilities that they do not yet possess. Human intelligence is more than just a heterogeneous set of skills. It is, first and foremost, the ability to understand the world and act within it to achieve goals, guided by an internal model that integrates experience, inference, and intention. This ability to integrate, rather than mastery of individual tasks, is its fundamental characteristic.

It is therefore surprising that these fundamental observations about AGI have escaped critical scrutiny, not only from representatives of technology companies, but also from high-level experts,

who deliver lengthy speeches and offer supposedly serious analyses of the likelihood of its imminent arrival and the dangers associated with it.

In conclusion, the vision of AI championed by tech giants—market-driven, obsessed with power plays, and mesmerized by its own myths—is not only misguided but also harmful. It has led to an extraordinary waste of intellectual and material resources. It is leading a remarkable scientific, technical, and intellectual adventure to a dead end. But dead ends invite us to reconsider, to choose another path—one that does not rely on scaling parameters or the pursuit of AGI. This path demands the difficult work of building systems we can truly trust, systems whose cognitive and deliberative capabilities we can progressively refine, until one day they might match our own.

5. From Leaders to Followers: The Academic Crisis

5.1 The Research Slop Crisis

Academic institutions have been caught up in the AI revolution without sufficient critical discernment. Eager to capitalize on funding opportunities, they have restructured priorities at the expense of traditional fields, leaving computer science in a deep crisis.

We are witnessing an explosion of technically empty papers following a standard template: numerous authors, dense citations, vague descriptions of concepts and problems without providing precise definitions, ad hoc experiments with intelligent prompts, anthropomorphic interpretations of results, and conclusions that align with mainstream optimism. This bears no resemblance to much of the rigorous computer science publishing of the past. It is, quite simply, a perversion of scientific methodology.

This has a direct impact on the evaluation and selection of papers submitted to conferences. The scale of the phenomenon is staggering: program committees must review more than 20,000 submissions, and AI reviewers—trained on existing literature—create a vicious cycle that weeds out genuine innovations. Many papers are poorly validated or entirely generated by AI; some contain completely fabricated information.
A recent Nature survey reveals that researchers are not blind to the risks—they embrace AI tools for their efficiency while fearing that these tools may encourage shallow analysis and irreproducible results [15].

Furthermore, the focus of AI research has shifted. Over the past decade, major tech companies have become the cornerstone of cutting-edge model development, while collaboration between academia and industry remains limited and is decreasing. The result is a dangerous disconnect: real progress occurs in industrial labs, while academic discourse focuses on constructing compelling narratives from those labs' results.

5.2 The Corporate Capture of Academic Research

The most worrying change is the reshaping of the research landscape itself. Research institutions have ceased to be drivers of knowledge production, ceding leadership to technology monopolies. They are becoming "junior partners"—subordinates who must follow corporate strategic choices. This is not merely a matter of commercialization but the privatization of an entire field's future: the questions worth asking, the paths worth exploring, even the definition of progress, are now set by a few corporations, not a diverse research community.

This weakening is evident in three trends:

The funding gap. US industry spent $340 billion on AI in 2021; the U.S. government invested $1.5 billion. Alphabet alone spent as much on DeepMind as the entire NSF computing research budget [16]. Academic researchers cannot access the infrastructure that industry labs take for granted, and corporate resources come with strings attached.

The talent drain. 70% of new US PhD graduates in AI now choose to work in the private sector, compared with 20% twenty years ago [17]. Young researchers are attracted by salaries, resources, and more interesting projects. Academia is left with a talent deficit, while the talent that leaves must conduct research within rigid, non-scientific frameworks that limit creativity.

Cognitive conformity. With limited resources, the academic world follows the guidelines set by corporations. The private sector dominates scientific discourse; to be published and gain recognition, academic work must align with prevailing viewpoints. Companies set the priorities; academia offers minor variations rather than new paradigms. Meanwhile, corporate labs themselves are moving away from openness, withholding technical details.

Academic research, once the main driver of innovative knowledge production, enabled the fundamental discoveries underlying modern AI. That role is now threatened. When research is judged by its applicability to corporate roadmaps, it loses the ability to ask questions that open new avenues of knowledge. This depletion of resources and misalignment of priorities foreshadows a future impoverished before it has even begun.

6. The Regulatory Vacuum: Why Governments Are Failing to Control AI

6.1 Restrictive vs. Permissive: The Atlantic Divide on AI Risk

For years, Western governments have watched the AI revolution unfold and let it happen, just as they did during the rise of the Internet. They celebrated innovation, funded research, and let technology companies operate with little or no oversight. In recent years, more and more voices have been raised to question the issue of regulation. Currently, official statements from governments and international institutions affirm the need for AI regulation. Yet there is no agreement on what can or should be regulated, nor on the types of risks involved. The gap between the European Union and the United States illustrates the depth of this disagreement.

The European Union has enacted the world's first comprehensive and enforceable AI regulation: the AI Act supplemented by the Digital Services Act [18]. These texts apply to AI systems the risk management approach long used for traditional artifacts. They define four risk categories for which they require a corresponding level of reliability. The strict application of EU texts under the current state of the art, will exclude many AI applications presenting an unacceptable or high risk, in particular critical autonomous systems.

In contrast, the US approach could not be more different. It has changed radically from one administration to the next. The Biden administration's executive order on AI (issued on October 30, 2023) [19], a voluntary and relatively permissive regulatory text, was rescinded by the Trump administration, which issued a new executive order stipulating self-regulation in the spirit of “innovation first” [20]. This text aligns perfectly with the industry's favorite narrative that standards are barriers to progress, as if progress were an end in itself, requiring no justification and allowing no restrictions.

Finally, it should be noted that, reportedly, the European Union faces difficulties in enforcing its legislation due to growing pressure from the United States to relax it.

Today, the chances of reaching agreement on a global regulatory framework, as the UN advocates, are currently slim. The United States is using its supremacy to impose a self-regulated, market-centric approach. Change will require counterbalancing this power through concerted action by countries that share the same values in order to harmonize standards for development and risk regulation.

6.2 The Imperative of Global Standards

The absence of global standards gives companies free rein to act in their own interests, creating de facto standards through their dominant market position rather than through an institution-controlled

process. Once established, these standards can prove irreversible. The industry's motto, “move fast and break things”, is actually a strategy that compromises the future: it leads to the construction of unreliable and vulnerable infrastructure that would be extremely costly to replace. We are already seeing this in the rush to deploy native AI systems. Too often, we worry about their safety and security after they have been deployed. This is a technically questionable approach, one that is likely to result in a spectacular collapse once these systems become operational and indispensable to the smooth running of the business. It creates *technical debt*, a concept with which software and systems engineers are well acquainted [21]. Its many dangerous manifestations—such as fatal failures or cyberattacks in traditional complex infrastructures—will become increasingly common as AI is gradually adopted.

The absence of global AI regulations creates a vacuum that powerful actors rush to fill. Without common standards, we get a race to the bottom in safety, the unchecked export of surveillance technologies, the permanent consolidation of corporate power, and the gradual erosion of democratic control over technologies that influence every aspect of our lives.

Global standards are the cornerstone of our technological civilization. We trust all kinds of artifacts, from toasters to airplanes, because they are built according to technical rules specific to their field.

To develop standards for AI, we need to agree on its well-known technical limitations and apply the precautionary principle. The argument that standards are an obstacle to innovation is completely unfounded. Applying standards will force us to conduct research and develop new solutions to overcome the identified limitations. Well-designed standards are essential to ensuring that innovation continues to serve society.

It is essential that future regulations adopt a holistic approach—one that understands and controls the benefits and risks of AI across all their dimensions. This requires three fundamental principles:

1. Reject reductionism. Regulatory frameworks must not reduce the problem to a simplistic trade-off between efficiency and safety, nor confine benefits to narrow metrics or risks to purely technical failures. The social, political, and human dimensions are equally consequential and must be addressed with the same rigor.

2. Account for AI's diversity. We must consider AI in its various forms, its different application areas, and its future development. Conversational systems, monitors, autonomous agents, and industrial controllers each raise distinct challenges and therefore demand tailored regulatory responses.

3. Pursue global consensus. AI is diffuse and borderless by nature. Its development, deployment, and consequences transcend national boundaries. No regulatory framework can succeed in isolation. Comprehensive solutions require international agreement grounded in shared values and mutual recognition of common stakes.

6.3 Three Faces of Risk: Technological, Anthropogenic, System

Discussions about the risks associated with AI lack clarity, particularly regarding the division of responsibilities among developers, operators, institutions, and users. A clear framework requires distinguishing three categories of risk.

Technological Risks arise when systems fail to perform as expected, endangering people or property. These engage the responsibility of developers who failed to take necessary precautions. For traditional systems, such precautions are specified by standards and regulations. For AI, the "black box" character makes evaluation exceptionally difficult. Specific technological risks include data-related (bias, poisoning), model-related (unpredictability, hallucinations, adversarial attacks), and security risks (inversion attacks, malicious exploitation). Detailed analyses are available in NIST documents [22].

Anthropogenic Risks stem from human activity, not machine failure. Two subtypes matter.

User-originated risks concern how people interact with AI systems. The fact that a language model generates fake content constitutes a technological risk; the fact that a user shares it is a matter of human action and responsibility.
When humans and machines must jointly perform critical tasks, the risks are poorly understood. This gap is clearly illustrated by the unresolved challenges associated with Level 3 autonomous driving. When the autopilot returns control to the driver, can we ensure that the driver understands the situation in time? When the driver takes control, can we guarantee a smooth transition? Current ad hoc protocols do not provide satisfactory answers to these questions.

Governance-originated risks stem from strategic decisions regarding the technical requirements of the system to be developed, as well as from communicating its capabilities and limitations to its users. Tesla's Full Self-Driving exemplifies the gap: it promises autonomy but requires supervision, creating mismatches between user expectations and actual performance [23]. Manufacturers exploit regulatory lag as competitive advantage, externalizing safety costs onto users and the public.

Systemic risks have the most far-reaching consequences, yet they are the ones we talk about the least. They stem from long-term transformations that are redefining social relationships, political institutions, and human cognition itself.

Some structural systemic risks are well known, are the subject of much debate, and have gained public attention. They include the concentration of power in the hands of technology monopolies, disruptions in the labor market, the erosion of the democratic deliberative process, and the rising environmental costs associated with large-scale computing.

Cognitive risks are deeper and more insidious systemic risks. They involve a transformation in the way humans think, brought about by the division of intellectual labor between humans and machines. Human and artificial intelligence are genuinely complementary. Machines excel at extracting knowledge from multidimensional data that go beyond human perception; humans possess reasoning and creativity machines cannot replicate. The danger is *cognitive offloading*—the tendency to use external tools to reduce mental effort. When offloading becomes habitual, the capacities being offloaded atrophy. Attention, memory, and reasoning weaken when we delegate their work.

This risk magnifies when AI enters education. Learning requires struggle—sustaining attention through difficult texts, wrestling with problems, constructing arguments through multiple drafts. AI tools that eliminate difficulty render training ineffective. We risk raising generations proficient with machines but lacking the competences machines offer them.

Three consequences follow:

Erosion of critical thinking. Accepting machine outputs without question atrophies the capacity to think slowly and with effort.
Weakening of responsibility. Delegating difficult choices to machines undermines our capacity for moral deliberation.
Homogenization of thought. AI systems trained on existing outputs produce the median, the expected. Humans consume and internalize these results, which become training data for future AI. Diversity of reasoning and expression converges toward stereotypes.

Paradoxically, these risks do not stem from failures, but from successes. The more capable AI becomes, the more urgently we face questions about the proper division of labor between humans and machines. It is time to address them before we lose the ability to do so.

7. The Silence of the Intellectuals

The lack of critical voices among intellectuals is another notable symptom of the AI era. During the first half of the 20th century, thinkers engaged in fierce public debate on the major issues of their time: totalitarianism, technology, freedom, and the nature of an ideal society. They were read, debated, and often despised, but they could not be ignored. They have stimulated social reflection and public debate.

Today, that function has largely collapsed. This is not because critical intellectuals no longer exist—they do, though in smaller numbers—but because they are no longer *audible*. The intellectual landscape is instead populated by a majority whose thinking aligns with dominant technological narratives, and a minority whose critical voices are systematically drowned out.

Where are the intellectuals who subject the fundamental assumptions of the AI era to the same level of scrutiny? Their absence from public view is the result of several converging trends discussed below..

7.1 Transformation of the Public Sphere: Why Critical Voices Are Not Audible

In the mid-20th century, intellectual debates were orchestrated by a small number of gatekeepers—editors-in-chief, publishers, and academic presses—who controlled the ideas reaching the public. Today, these gatekeepers have disappeared. It is much easier to publish thanks to direct access to the media, at virtually no cost.

But this openness has created a new problem: the public square has been replaced by an algorithmic feed that ignores criteria of quality and relevance; it amplifies anything that sparks engagement, outrage, wonder, sensationalism. Nuanced and critical voices are not censored, but they simply do not find the resonance they deserve and are drowned out by the background noise.

The result is paradoxical: more voices than ever, but less capacity to hear the ones that matter. Critical thinkers still exist, but their words vanish while the superficial rise. It is far more rewarding to be an apologist, a famous “futurist” who validates the existing power structure, than a critical theorist who questions its very foundations. Riding the wave of powerful technological narratives—the coming singularity, the revolutionary potential of blockchain, the Metaverse, the inevitable march of AGI—is a shortcut to visibility.

7.2 The Mechanisms of Capture

Over the past two decades, dominant technology players have built an extensive apparatus of influence by funding non-profits, working groups, and think tanks. These organizations appear independent but their funding originates from tech foundations, and their programs remain subtly aligned with corporate interests.

The result is that the entire intellectual ecosystem is permeated by industry influence. The questions worth asking, the methods worth using, the conclusions worth reaching—all are subtly crafted by corporate power, producing a simulacrum of independent intellectual life that serves its interests.

This institutional capture is reinforced by ideological capture—pseudo-philosophical frameworks that legitimize corporate dominance.

The singularity narrative—a cosmic eschatology promising technological transcendence—portrays machine intelligence as an unstoppable force toward a predetermined future, rendering any opposition narrow-minded [24]. This intoxicating narrative serves technology companies perfectly.

Technological determinism presents technology as an autonomous force to which society must simply adapt [25]. This anti-humanist ideology denies humans any role as actors in their own

evolution. Intellectuals struggle to explain to the public how to adapt and “cope” with the inevitable rather than challenge it.

The message is clear: companies do not make choices; they are merely vectors of inevitable progress. They cannot be blamed for job losses or social disruption—these are just "progress." This ideology absolves them of moral and political responsibility while discouraging citizens from reacting.

7.3 Asymmetric intimidation

It is often claimed that intellectuals are intimidated by AI's technical complexity, fearing that any factual error will expose them to viral online attacks. There is some truth to this, but this explanation is incomplete. If technical ignorance were truly disqualifying, we would not see a proliferation of pseudo-thinkers whose superficial understanding of AI is treated with unwavering seriousness. Those who make the most implausible claims—about AI sentience, imminent superintelligence, or human obsolescence—are amplified by the media, celebrated as visionaries, and rewarded with attention.

The paradox is stark: a cautious critique must be flawless to survive, while grandiose absurdity faces no consequences. The system does not punish falsehood; it rewards it, provided it aligns with the dominant discourse. Pseudo-thinkers thrive because their narratives are emotionally satisfying and ideologically useful. They are not intimidated by technical challenges and complexity—quite simply, they don't care. Their sole concern is reaching the widest possible audience.

7.4 The Abdication

The silence of intellectuals is not accidental. It is the product of three converging forces: an attention economy that amplifies sensation over substance, an institutional landscape permeated by corporate power, and a media environment that rewards absurdity while holding critical thought to impossible standards. These forces ensure that the voice of the critical minority is not heard.
What is lost is an essential social function—the rich, contradictory debates that once stimulated critical thinking and raised uncomfortable questions. By failing to break through the noise, or by joining the chorus of apologists, intellectuals leave the field open to the very forces that should be the subject of their critical scrutiny.

8. Facing Soft Obscurantism: Between Acquiescence and Action

The age of AI brings radical changes: private control of infrastructure, the domestication of academic thought, and a culture that blurs the lines between reality and fiction—all of which demand our submission to corporate power.

Humanity has already overcome obscurantism in the past. The Middle Ages came to an end thanks to collective action and the establishment of new institutions. But this crisis differs in two essential ways. First, speed: centuries back then, years today—societal antibodies don’t have time to develop. Second, control: the authoritarian regimes of the past relied on external force; today’s domination operates subliminally, almost consensually, shaping cognition itself.

Our inertia has three sources: convenience (immediate pleasure, deferred costs), complexity (real difficulty and muddled discourse), and ideology (corporate visions that portray opposition as a rejection of progress). Traditional checks and balances have weakened.

Yet this decline is not irreversible. Breaking the deadlock begins with naming the problem and revitalizing public intellectual debate. The issues raised by AI are too important to be left to technologists and business leaders. We need new counterweights to private power— grassroots movements, civic alliances, and public platforms that can challenge corporate dominance.

The dominant discourse presents AI as either an unstoppable force or an existential threat. These two narratives lead us to debate machines rather than confront human failings. Silent acceptance is the first act of a larger drama. Neither fate nor demon. The problem is not AI—it is our unwillingness to manage it rationally and responsibly.

Note: This paper includes, among other things, a personal technical analysis, which will be the subject of future publications. I have primarily cited sources containing verified facts, as I did not feel it was appropriate to burden the text with an excessive number of technical references. AI tools were used solely to facilitate the writing process and the collection of information.

References

[1] Joseph Sifakis and David Harel (2023). Trustworthy Autonomous System Development. ACM Transactions on Embedded Computing Systems, Volume 22, Issue 3**,** Article No.: 40, Pages 1 – 24 https://doi.org/10.1145/3545178

[2] CBS News. (2019, April 22). Elon Musk predicts Tesla will have "robotaxis" on the road next year. https://www.cbsnews.com/news/elon-musk-claims-tesla-will-have-self-driving-robotaxis-on-the-road-next-year/

[3] Atento. (2025, December 10). *Four AI and CX predictions that didn't come true this year – and what we learned*. https://atento.com/en/insight/four-ai-and-cx-predictions-that-didnt-come-true-this-year-and-what-we-learned

[4] Lenze. (2025). *SIL - Safety Integrity Level*. Lenze Application Knowledge Base. https://www.lenze.com/en-lt/application-knowledge-base/article/282055/1

[5] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[6] Koopman, P., & Widen, W. (2024). *Redefining Safety for Autonomous Vehicles*. SafeComp 2024. arXiv:2404.16768. https://arxiv.org/abs/2404.16768

[7] Digital Watch Observatory. (2025, September 27). *Digital on Day 4 of UNGA80: Governance, inclusion, and child safety in the AI age*. https://dig.watch/newsletters/events/digital-on-day-3-of-unga80-2

[8] Amol S. Dhaigude and Giridhar B. Kamath (2025). *Mapping responsible artificial intelligence in business and management: Trends, influence, and emerging research directions*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2199853125001751

[9] Momentè, F., Suglia, A., Giulianelli, M., Ferrari, A., Koller, A., Lemon, O., Schlangen, D., Fernández, R., & Bernardi, R. (2025). *Triangulating LLM Progress through Benchmarks, Games, and Cognitive Tests*. University of Edinburgh. arXiv:2502.14359. https://arxiv.org/abs/2502.14359

[10] The Register. (2026, January 30). *Deciphering the alphabet soup of agentic AI protocols*. https://www.theregister.com/2026/01/30/agnetic_ai_protocols_mcp_utcp_a2a_etc/

[11] Sutskever, I. (2025, November). Interview with Dwarkesh Patel. In *The Dwarkesh Podcast*. "Ilya Sutskever: The End of the Scaling Era." https://www.dwarkesh.com/p/ilya-sutskever-2

[12] Heath, A. (2025, August 1). Zuckerberg's 'personal superintelligence' plan: fill your free time with more AI. The Verge. https://www.theverge.com/command-line-newsletter/717880/zuckerbergs-personal-superintelligence-plan-ai-chatgpt-race

[13] Gartner. (2025, June 25). *Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027* [Press release]. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

[14] Coldewey, D. (2025, May 25). *From LLMs to hallucinations, here's a simple guide to common AI terms*. TechCrunch. https://techcrunch.com/2025/05/25/from-llms-to-hallucinations-heres-a-simple-guide-to-common-ai-terms/

[15] Van Noorden, R., & Perkel, J. M. (2023). *AI and science: what 1,600 researchers think*. Nature, 621, 672-675. https://www.nature.com/articles/d41586-023-02980-0

[16] (Ahmed, N., Wahed, M., & Thompson, N. C. (2023). *The growing influence of industry in AI research*. Science, 379(6635), 884-886. https://www.science.org/doi/10.1126/science.ade2420

[17] Stanford University Human-Centered Artificial Intelligence. (2024). *Artificial Intelligence Index Report 2024—Education Chapter*. https://hai.stanford.edu/ai-index/2024-ai-index-report/

[18] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689, 12 July 2024, pp. 1-144. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[19] The White House. (2023, October 30). *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Executive Order 14110, 88 FR 75191. https://www.govinfo.gov/content/pkg/FR-2023-11-01/html/2023-24283.htm

[20] The White House. (2025, December 11). *Ensuring a National Policy Framework for Artificial Intelligence*. Executive Order 14365. https://www.whitehouse.gov/fact-sheets/2025/12/fact-sheet-president-donald-j-trump-ensures-a-national-policy-framework-for-artificial-intelligence/

[21] Fowler, M. (2006, October 13). *Technical Debt.* https://martinfowler.com/bliki/TechnicalDebt.html

[22] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1.

[23] National Highway Traffic Safety Administration. (2023). *ODI Investigation: Tesla Full Self-Driving*. NHTSA Action Number: PE23020. https://static.nhtsa.gov/odi/inv/2022/INCR-EA22002-14496.pdf

[24] Kurzweil, R. (2005). *The Singularity Is Near: When Humans Transcend Biology*. Viking Press.

[25] Brayford, K. (2020). Myth and technology: Finding philosophy's role in technological change. Human Affairs, 30(4), 526-534. https://doi.org/10.1515/humaff-2020-0045