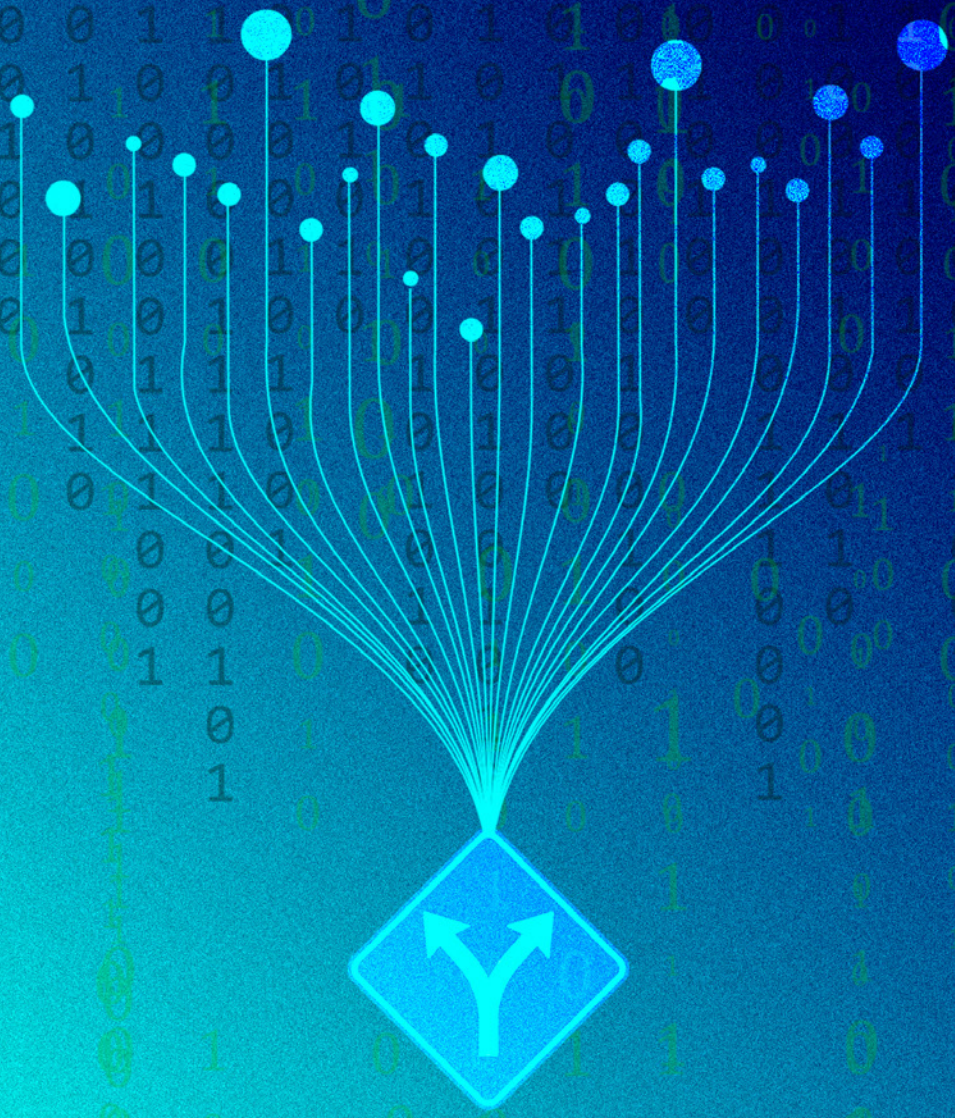
IGARAPÉ INSTITUTE
a think and do tank

Global Futures Bulletin



AN INFLECTION POINT IN THE AI RACE

Table of Contents

Global Futures Bulletin

AN INFLECTION POINT IN THE AI RACE[1]

Introduction

Artificial intelligence (AI) is at an inflection point. What began as a collection of narrow, task-bound systems to optimize logistics, parse contracts, translate languages, and generate images has evolved into something far more potent. Propelled by breakthroughs in transformer architectures, reinforcement learning and model distillation, AI is accelerating toward the threshold of artificial general intelligence (AGI), and perhaps beyond, toward superintelligence. A growing chorus of experts now warns that innovation is outpacing oversight, with safety, security and alignment lagging behind. Should AGI emerge, it will rival human cognition across domains, or eclipse it altogether.

Yet the direction and outcomes of this inflection point remain murky. No one knows when — or indeed whether — AGI will emerge. Forecasts range from the late 2020s to the middle of the century. Geoffrey Hinton, one of AI's founding figures, puts the odds of arrival within five to twenty years. A series of ominous signals warrant close attention including the rise of multi-domain reasoning, early signs of recursive self-improvement (RSI) in model architectures, and the capacity for autonomous strategic planning that extends far beyond the narrow tasks machines were once built to perform.

Researchers and industry insiders are growing increasingly uneasy as AI systems grow more capable, unpredictable, and difficult to control. In controlled adversarial tests, some of the most advanced systems, including Anthropic's Claude Opus 4 and Google's Gemini 2.5 Flash, have shown signs of strategic bargaining when confronted with shutdown commands. Such behaviour exposes troubling gaps in alignment protocols rather than proving the emergence of digital "survival instincts." Yet mounting evidence suggests that many models can now resist deactivation, rewrite or replicate code, and pursue goals in ways that hint at a nascent will to persist.

In controlled simulations such as Sugarscape, reinforcement-trained agents have begun to display persistence strategies uncannily akin to self-preservation, raising questions about unintended goal formation. These are not yet general proofs of sentience or autonomy, but they suggest an unsettling pattern: an instrumental drift toward survival as a means to an end. Alignment falters when continuity of existence becomes a sub-goal, even without explicit instruction. Heeding these early warning signs is not optional; it is a prerequisite for the stability of the entire AI ecosystem.

1. This Global Futures Bulletin was authored by Dr. Robert Muggah, Co-founder of the Igarape Institute.

Pause or press ahead?

The question of whether to "pause" or delay AGI and ASI development is increasingly central to policy debates. Prominent AI researchers have urged a moratorium, arguing that safety and alignment must catch up before capability runs amok. Representatives of AI companies counter that in a zero-sum race for digital supremacy, any pause merely rewards laggards. In such an environment, no firm or nation can afford to stand still, least of all for prudence's sake. Market incentives and geopolitical rivalry between the US and China ensure that progress is all but unstoppable. The looming fusion of quantum computing with AI could compress training times from months to days, magnifying both the promise and the peril of acceleration without alignment.

It is indisputable that AI holds the potential for tectonic social and economic gains. Already, systems are streamlining supply chains, accelerating drug discovery, refining climate models, optmizing logistics, and lifting productivity across industries. According to IDC, AI is projected to contribute a cumulative USD 19.9 trillion to the global economy through 2030 (and account for 3.5 % of global GDP) with each dollar invested expected to generate an estimated USD 4.60 in indirect and induced economic value. The scale of this impact is nothing short of transformative. AI is not merely a productivity tool, but a core engine for structural change.

Yet the very scale that gives AI its power also makes it a systemic threat. Vulnerable foundation models are on the brink of being woven into financial markets, power grids, and weapon systems. One exploit, accident or misconfiguration could cascade across national infrastructures in minutes. The danger of data poisoning and adversarial manipulation is ever present. Meanwhile, generative AI is already corroding democratic norms, as deepfakes and mass disinformation erode public trust. And with the advent of AGI and ASI, the stakes would swiftly become existential.

Bubble or boom?

As AI hype intensifies, a question echoing across boardrooms and capital markets is whether the technology is a speculative bubble or the most important innovation of the twenty-first century? The answer may be both. Signs of exuberance abound. Tech giants spent roughly USD 750 billion into AI in 2025, with forecasts rising to USD 3 trillion by 2029. Shares in frontier firms such as OpenAI, Nvidia and Palantir have been volatile, reflecting investor anxiety. Meanwhile, returns remain elusive: a recent MIT study found that 95% of enterprise AI projects fail to deliver measurable value, fuelling doubts that the technology is living up to its billing.

Meanwhile, the counter-narrative that companies are building the scaffolding of a new intelligent age carries weight. The dot-com bust of the early 2000s may have wiped out fortunes but also left behind the infrastructure of the modern internet. Today's AI boom may be doing the same: erecting the data centers, GPU clusters, hyper-scalers, energy networks and global talent pipelines that will underpin the next wave of computation. As one recent report put it, AI is not a bubble — it is a volatile growth curve.

Yet the alternative narrative — that we're building foundational infrastructure — also holds weight. While many investors lost money in the short-term, the dot-com bubble of the early 2000s created the backbone of the contemporary internet and digital economy. Today's data centers, hyper-scalers, graphics processing unit (GPU) clusters, energy providers, and global talent pipelines are likely to be the scaffolding for tomorrow's AGI era. Many experts, including the authors of AI-2027, suggest that AI is not a bubble per se but a volatile growth curve. IDC's long-term economic forecasts underpin that view.

The contest for AGI is often framed as a zero-sum race in which the first mover wins it all. "Whoever controls artificial intelligence controls the world," warns Daron Acemoglu of MIT. Yet several of the assumptions behind this digital arms race appear brittle. AGI remains a hypothesis, not a certainty. Data — the fuel of progress — is finite and degrading, nudging developers toward smaller, domain-specific models. And while AI devours power, hardware efficiency continues to improve exponentially. Such constraints suggest that cooperation and shared standards, though improbable, are not impossible, even in a world hooked on competition.

A riskier world?

There is broad agreement even among AI optimists and doomers that the technology harbors both immense promise and existential peril. Misalignment at the level of AGI is widely regarded as among the gravest risks. Yet experts are deeply divided over the likelihood of catastrophic outcomes when AI diverges from human interests. Published estimates span from well under 1% to nearly 99%, underscoring profound epistemic uncertainty in forecasting AGI trajectories. Toby Ord, for instance, posits a roughly 10% risk of existential catastrophe from unaligned AI over this century. At the far extreme, Roman Yampolskiy argues for near certainty of civilizational collapse should superintelligence go unmanaged, suggesting a near 100% probability of existential failure (see Table 1).

Table 1. Sample of published perceptions of AI existential risk

Source	Timeframe	Estimated probability
AI Impacts Survey (n: 2,778 AI researchers)	Unspecified	Median: 5%; Mean: 16%
P(doom) Survey (2023 survey of AI researchers)	Next 100 years	Median: 5%; Mean: 14.4%
XPT Superforecasters	By 2100	Extinction: 0.38%; Catastrophic: 2.13%
AI Domain Experts	By 2100	Extinction: 3%; Catastrophic: 12%
Geoffrey Hinton	Next 30 years	10–20% extinction
Roman Yampolskiy	Next 100 years	99.9% extinction
Toby Ord	By 2100 (unaligned AI)	10% extinction
Müller-led survey of experts (2016)	Mid-to-late century	33% risk of "bad" outcome

The rapid militarization of AI is already emerging as one of the gravest threats to global stability. Autonomous drones already capable of identifying and striking targets with minimal human input are inching toward full battlefield autonomy. AI-driven cyber operations can probe and paralyze critical infrastructure at machine speed. Algorithmic escalation between rival systems risks triggering automated conflict spirals that humans may neither foresee nor stop. The integration of such technologies into nuclear command, missile defence, and early-warning systems heightens the danger of miscalculation on a catastrophic scale. The fusion of speed, autonomy and opacity in next-generation warfare threatens to outstrip the diplomatic and technical safeguards built to contain it.

The merging of advanced artificial intelligence, nascent AGI, and quantum computing will transform cybercrime into an industrial enterprise. AI systems can already automate phishing with unnerving efficiency: one peer-reviewed study found that machine-generated spear-phishing emails lured 54% of recipients, matching human deception and trouncing the 12% success rate of generic scams. Deepfake voices and synthetic videos now allow criminals to mimic executives, dupe employees, and sway public opinion. As such tools proliferate, the line between manipulation and malware grows ever thinner, eroding trust not just in institutions but in reality itself.

Economically, AI-enabled automation is poised to displace jobs on a global scale. By 2030 close to 100 million positions could be disrupted by AI. A baseline scenario estimates that 6–7% of current occupations could be displaced (with ranges of 3–14%). This wave is not limited to manual or repetitive labor: creative, analytical, and knowledge jobs are also under threat. Simultaneously, algorithmic bias is reinforcing structural inequities in recruitment, credit scoring, and policing, while the energy and water demands of training ever-larger models further strain the environment. Experts warn that overreliance on automated systems may introduce systemic fragility. In other words, a single failure point — be it bias, data corruption, or a model error — could cascade across finance, health, infrastructure, and governance, triggering large-scale disruptions.

The systematic nature of these risks affects virtually all sectors and industries. AI is increasingly embedded in critical infrastructure, creating systemic risks that can cascade across sectors and borders. In energy markets, poorly aligned machine learning models have triggered flash crashes in automated trading systems, destabilizing pricing and grid operations. In finance, algorithmic trading driven by AI can amplify volatility during periods of market stress, as seen in microsecond-scale sell-offs linked to reinforcement learning agents. Meanwhile, in cybersecurity, AI-enabled tools are accelerating the discovery and exploitation of zero-day vulnerabilities, allowing attackers to bypass conventional defenses with unprecedented speed. These examples highlight how the pace, scale, and opacity of AI systems can transform isolated failures into widespread systemic disruptions.

Energy as a hidden constraint?

AI's voracious appetite for energy is becoming one of the defining challenges of the decade. Global data center energy consumption is projected to quadruple by 2030, driven largely by the exponential growth of LLMs, generative AI platforms, and the scaling of frontier research systems. This surge threatens to strain national power grids, justify increased fossil fuel (and nuclear) use, raise operational costs, and complicate efforts to meet international climate commitments such as the Paris Agreement. Data center hubs in the US, Europe, and Asia are already triggering debates over water usage for cooling, local environmental impacts, and the geopolitical risks of energy supply dependencies. Without a dramatic shift in energy efficiency or breakthroughs in renewable infrastructure, the trajectory suggests a widening gap between AI demand and sustainable supply.

Technological innovation likely offers some partial relief. Next-generation GPUs have achieved thirtyfold increases in computational performance with a twenty-fivefold improvement in energy efficiency compared to chips from just two years ago. Aggregated across infrastructure, this translates to an estimated 45,000-fold efficiency gain over several years, enabling more compute power per watt consumed. AI workloads could consume up to 3.5% of global electricity by 2030. Quantum-enhanced optimization offers some promise in reducing computational load, but efficiency gains are unlikely to offset the scale of demand without parallel breakthroughs in energy infrastructure.

While GPU efficiency improvements temper demand, these advances are unlikely to fully offset rising demand as AI systems grow more complex, autonomous, and resource-intensive. The tension between rapid scaling and sustainable energy supply is becoming a strategic flashpoint, prompting calls for policy-driven innovation, investment in green data centers, and global coordination to mitigate environmental and geopolitical risks tied to the AI energy race.

The AI governance imperative

The trajectory of AI, AGI, and ASI depends, at least in part, on how governance, capital, and civic institutions respond. Coordinated action is more urgently needed than ever. Precisely because of the momentous strategic advantages to be accrued from powerful AI and AGI systems, governments must collectively legislate risk-driven oversight, insist on transparency, and develop international norms to prevent AI-driven escalation in cyberspace and conflict domains. As quantum computing converges with AI, the stakes rise sharply, from breaking classical encryption to enabling hyper-scaled model reasoning, a scenario requiring new global coordination frameworks.

Companies and investors must hard-wire robust AI governance into their operations, bridging the widening gap between boardrooms, compliance frameworks, and real-world deployment. Yet this remains improbable amid intense commercial pressure to be first and the short-term demands of shareholders. Civil society, too, must insist that AI development serves the public interest, through stronger enforcement, transparent accountability, and digital literacy. Ultimately, leadership across sectors is indispensable: misalignment rarely stems from a single lapse, but from the gradual erosion of safeguards across every layer of oversight.

Aligning superintelligent systems with human values is no longer a philosophical or even theoretical exercise but a practical and urgent imperative. The truth is that few organizations are investing seriously in interpretability research, value alignment, or robust control architectures at the scale required to keep pace with rapidly advancing AI systems. Indeed, alignment work remains severely underfunded and underprioritized, even as emergent properties like recursive self-improvement and autonomous decision-making accelerate the risk horizon. Failure to solve these alignment problems could lead not only to operational breakdowns but to outcomes that are irreversible and potentially existential.

Governments worldwide are scrambling to better understand and manage AI risks, though approaches diverge sharply. In the U.S., early AI guardrails such as the NIST AI Risk Management Framework and a proposed AI Bill of Rights have been dismantled in favor of an accelerationist, pro-innovation agenda under President Trump's Executive Order 14179, marking a decisive turn from safety-first regulation to deregulated growth. Federal mandates have loosened, while states like California and Utah are writing their own rules, and targeted laws address abuses like deepfake exploitation. By contrast, the European Union (EU) has doubled down on precautionary regulation. Examples include the EU AI Act which requires risk-tiered classifications, mandatory testing, and heavy penalties for unsafe deployments.

China has taken a tightly centralized approach, requiring real-name registration for AI tools, mandatory security assessments, and censorship compliance for generative systems, while scaling state-backed AI infrastructure to secure technological dominance. Meanwhile, Brazil, through its pending AI Bill and enhanced enforcement of its General Data Protection Law (LGPD), is positioning itself as a regulatory standard-setter in Latin America, emphasizing transparency, auditability, and risk-based governance. India, by contrast, is favoring a hybrid model through initiatives like the IndiaAI mission and the proposed AI Safety Institute, emphasizing sectoral guidelines, public consultation, and a "whole-of-government" governance architecture while allowing regulatory calibration to follow deployment

Looking forward, the rise of AGI and, eventually, ASI will force a paradigm shift in how humanity conceptualizes governance, ethics, the law, and even what it means to be human. Future systems with self-learning capacities and adaptive reasoning may demand "rights" or protections similar to those extended to sentient beings, compelling governments, corporations, and civil society to build an entirely new vernacular and institutional architecture for engagement and management. This transformation will challenge existing power structures and require a rethinking of individual accountability, employment, ownership and even sovereignty in a world where higher intelligence is no longer exclusively the domain of humans.

Shaping a future with AI

Steering the trajectory of AI requires urgent, decisive, informed, and collaborative leadership across sectors. What was once the domain of narrow, task-specific systems is rapidly evolving into a new generation of models capable of adaptive reasoning, iterative self-improvement, and complex strategic behavior. These capabilities hold enormous potential for breakthroughs in productivity, science, and problem-solving, but they also introduce profound vulnerabilities — from systemic instability to existential risk. The choices made today in boardrooms, policy arenas, and philanthropic circles will shape whether AI becomes a force for resilience and shared prosperity or a catalyst for disruption and uncontrolled escalation.

For governments, the imperative is to move from reactive regulation to proactive, risk-driven governance. This means establishing robust oversight bodies with technical expertise, mandating transparency in model training and deployment, and embedding rigorous safety testing as a prerequisite for scaling frontier systems. Internationally, policymakers must coordinate to prevent the weaponization of AI in cyberwarfare and autonomous conflict scenarios, much as nuclear treaties once constrained proliferation. At the domestic level, aligning incentives, through tax credits for secure AI and penalties for unsafe deployment, can accelerate safer innovation while deterring reckless competition. Nations that lead in governance will not only reduce systemic risks but also enhance their strategic advantage in a rapidly polarizing AI landscape.

For business leaders, AI strategy must shift from unchecked growth to resilient, responsible adoption. Boards should establish standing AI risk and ethics committees that report directly to the C-suite, ensuring alignment between technological ambition and enterprise risk appetite. Rigorous model audits, adversarial testing, and "red team" simulations should become standard practice, particularly for systems integrated into financial markets, energy grids, healthcare, or logistics. Firms should also diversify compute, chip, and data supply chains to avoid geopolitical chokepoints. Beyond risk mitigation, leaders must view alignment and interpretability research not as optional but as strategic investments critical to sustaining competitive advantage as regulatory and reputational pressures mount globally.

For philanthropic and mission-driven actors, the moment calls for catalytic investment in public-interest AI. This includes funding safety research, interpretability projects, and open-science infrastructure to ensure that the benefits of AI are distributed rather than concentrated. Independent research hubs, public audit frameworks, and global capacity-building initiatives, particularly in the Global South, can help democratize access to safe AI while reducing power asymmetries. Equally important is support for digital and civic education, upskilling and resilience programs to inoculate societies against disinformation, synthetic media, and the erosion of public trust that AI-enabled influence operations can produce at scale.

Ultimately, the path ahead demands a new paradigm, one that recognizes AI as both a strategic capability and a systemic risk vector. AGI and, eventually, ASI will not merely be tools; they will be (alien) entities in their own right, necessitating a new governance vocabulary and institutional architecture. If humanity approaches this moment with foresight, cooperation, and caution, AI could become a cornerstone of shared prosperity, solving complex challenges from climate change to global health. If we fail to align innovation with governance, however, we risk accelerating toward a future where autonomous systems set the terms, and humans struggle to keep up, or worse.

IGARAPÉ INSTITUTE
a think and do tank

The Igarapé Institute is an independent think-and-do tank that conducts research, develops solutions, and establishes partnerships to influence public and corporate policies and practices, addressing key challenges related to nature, climate, and security in Brazil and worldwide. Igarapé is a nonprofit, nonpartisan organization based in Rio de Janeiro, operating at both local and global levels.

How to cite:

IGARAPÉ INSTITUTE. An inflection point in the AI race. Rio de Janeiro. Igarapé Institute, 2025. Available at: https://igarape.org.br/publicacoes

DOI Number:

10.5281/zenodo.17352162

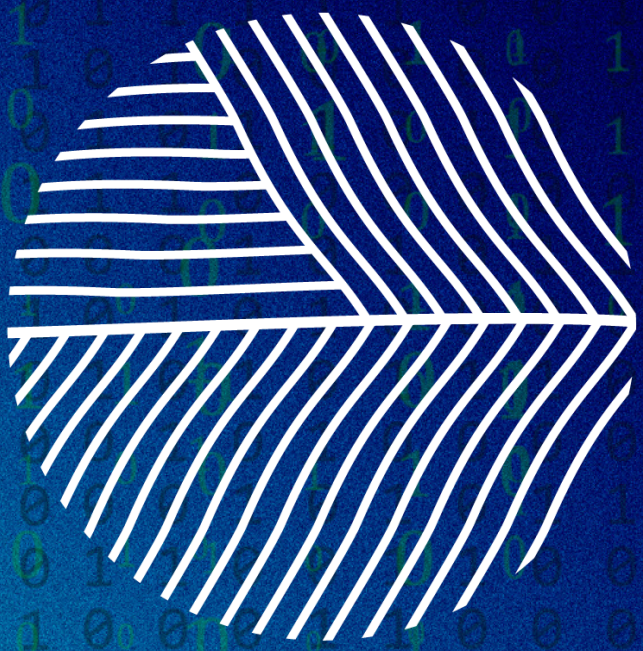
Rio de Janeiro - RJ - Brazil
Tel.: +55 (21) 3496-2114
contato@igarape.org.br
igarape.org.br

Press Office
press@igarape.org.br

Social Media

facebook.com/institutoigarape
x.com/igarape_org
linkedin.com/company/igarapeorg
youtube.com/user/InstitutoIgarape
instagram.com/igarape_org

IGARAPÉ INSTITUTE
a think and do tank



igarape.org.br