Highlights of the Issue

How We're Using AI to Produce *SuperIntelligence*

We're experimenting with AI in three ways to save your time and keep you informed about cutting-edge AGI developments:

1. Producing summaries of significant articles instead of providing full texts (testing various LLM models)
2. Producing critical reviews of articles that we do have space to include or are under review elsewhere.
3. Using the two DeepResearch models by OpenAI and Google to produce in-depth reports on important topics, especially lengthy ones.

Regarding the critical reviews, while LLMs are more knowledgeable than any one subject-matter expert in most subjects, their critique of a given paper may not be as astute as the expert's. Nonetheless, as LLMs close this gap, they are much more efficient than the peer-review process and, for most of the articles we publish, more than adequate. Further, at this point in AGI development, we prefer public debate over traditional opaque peer-review.

Your comments and suggestions on this effort are welcome.

Editorial: The Perilous State of AI Governance, June 2025

There are over 1000 AI governance bills pending at the federal and state level in the US. To avoid an inefficient 'patchwork' regulatory regime, a federal bill (relevant text provided in the editorial) calls for a 10-year moratorium on state AI regulation. A California bill proposes competing, private regulatory bodies licensed by the Attorney General. We support these efforts.

The editorial image of the US flag and skull was created with Copilot.

The First International AI Safety Report

> *Overall, the risk assessments in this report should be read with the understanding that AI has gained capabilities since the report was written.*
> *– Yoshua Bengio*

Chaired by Yoshua Bengio, a group of 96 researchers from 30 countries collaborated on this comprehensive report with 1366 references. Expect a thorough overview of all aspects of AGI safety, including categories of solutions ('risk mitigation'), but don't expect actual detailed solutions. It is more of a complete guidebook for workers who want to get the big picture of the AGI safety paradigm. For specific AGI safety solutions, see, e.g., the pragmatic Uuk et al. "Effective Mitigations for Systemic Risks from General-Purpose AI," and the theoretical Bengio et al., "Can a Bayesian Oracle Prevent Harm from an Agent" – both in our previous issue – and our other articles.

Of the ~300-page report, we excerpt 50 pages:

- Abstract
- Executive Summary
- Chair's Note: Update on the latest AI advances after the writing of the Report

- Key Findings of the Report
- 3.2 General challenges for risk management and policymaking
- 3.4 Risk mitigation and monitoring

We also include a 42-page Review of the entire report in the Reviews section. For the references, go to the full report.

An excerpt from Risk mitigation and monitoring:

> *The main evidence gaps around risk management for general-purpose AI include how great the risks are, and the degree to which different mechanisms can actually constrain and mitigate risks in real-world contexts. There is not always a scientific consensus about how likely or severe the risks of general-purpose AI systems are or will be, making it difficult for policymakers to know whether and how they should be prioritised. For example, how to manage misuse risk will depend on how skilled threat actors are in real-world contexts of concern. Moreover, most of the risk management efforts described above are not yet validated, standardised, or widely used. Risk management efforts vary across leading AI companies and incentives may not be well-aligned to encourage thorough assessment and management (982). While there are a few risk mitigations that are perceived to be the most effective by experts for reducing systemic risks from general-purpose AI (983), the efficacy of general-purpose AI risk management mechanisms is still being assessed and policymakers should seek more evidence from real-world applications.*

A Framework for the Private Governance of Artificial Intelligence

Dean Ball, coming from a classical liberal background (with which this editor strongly sympathizes), proposes a hybrid public-private AGI governance structure. Governing bodies at the state/province, federal, or international level would license private regulators to compete in monitoring AGI makers for compliance with existing safety standards. AGI makers are not required to be licensed by the private regulatory bodies but are incentivized by being indemnified from tort liability by the governing body of law if they are in compliance with the regulatory bodies' requirements.

In principle, Ball's hybrid approach has merits, but serious problems remain to be solved, which I think they can be. For one, competition between the regulatory agencies incentivizes them to provide thorough, effective regulation in a highly dynamic AGI development environment, as judged by the governing body granting their license. However, I see no mechanism to disincentivize AGI providers from seeking the least-onerous regulator.

Second, Ball suggests that the private regulators audit AI providers once per year. This is far too infrequent in the AGI arms race. Mechanisms need to be constructed so that the private regulatory bodies are not incentivized to keep costs down by regulating at too-infrequent intervals, and at the bureaucratic State level, to review the private regulators at too-infrequent intervals. On the other hand,

I think auditing the AGI providers *at least* once per month is necessary. Better still, automating as much of the audit as possible and preventing hacks on the audit programs and audit trail is highly desirable, especially in preparation for a hard takeoff.[1]

California Senate Bill 813: A Novel Approach to Artificial Intelligence Governance

Sponsored by State Senator McInerney, this bill embodies the hybrid public-private governance structure advocated by Dean Ball in the preceding article. The Attorney General would license private, competing 'multistakeholder' organizations to regulate AI companies. Companies would be incentivized to be certified by the regulatory organizations to reduce their potential legal liability.

LLM Security: Vulnerabilities, Attacks, Defenses, and Countermeasures

> *This survey seeks to define and categorize the various attacks targeting LLMs, distinguishing between those that occur during the training phase and those that affect already trained models. A thorough analysis of these attacks is presented, alongside an exploration of defense mechanisms designed to mitigate such threats.*

AutoRedTeamer: Autonomous Red Teaming with Lifelong Attack Integration

This is a preprint undergoing review at a refereed journal.

As we move toward artificial general intelligence, safety technology must shift from human supervision to automated routines. Further, some safety routines will have to run in real-time. "AutoRedTeamer combines a multi-agent architecture with a memory-guided attack selection mechanism to enable continuous discovery and integration of new attack vectors."

Pitfalls of Evidence-Based AI Policy

As scientists, our default belief is all policy should be "evidence-based." What if the evidence fails to capture significant risks due to the novelty and fast-evolving nature of the target technology? What if the evidence is biased? What if excessively high evidentiary goals delay putting regulation in place? Casper et al. bore into these and related issues at this critical juncture in AI safety and alignment regulatory development.

Strategic Patience: Long-Horizon AI Dominance and the Erosion of Human Vigilance

Yampolskiy has systematically tried to cover all aspects of AGI risks and the weaknesses of safety and value alignment responses. Here he looks at scenarios where overt AGI is not imminent but plays a long game to assure its dominance over human control. Humans become increasingly dependent on AI, continuing the current trend and lulled into complacency regarding AGI risks. Policy implications in this scenario are examined.

---

[1] Carlson, K. W. (2019). Safe artificial general intelligence via distributed ledger technology. *Big Data Cogn. Comput., 3*(40). doi:10.3390/bdcc3030040.

Quantum Immortality

Sean Carroll persuasively argues that 'many worlds' is not an *interpretation* of current quantum theory, but an inherent component of the theory itself.[2] An implication of many-worlds is that in some of those worlds an individual human, or the entire species, will survive ('quantum immortality') even if in other worlds, they perish.[3] Here, Turchin and Miller speculate on the implications of assuming high AI doom probability and quantum immortality.

> *The idea of quantum immortality is that the fate of an observer from their own point of view differs from the fate of a random mind: the observer will survive. What we described here adds to this idea: the observer will not only survive but will be more likely to exist in a world where AI Doom will not happen. For the same reasons, other catastrophes will also not happen in most of my surviving futures.*

In the audiobook cited, Sean Carroll comments thoughtfully and thoroughly on "Free Will, Determinism, and Many Worlds" (Lecture 22) and "What Happens to Ethics under Many-Worlds," (Lecture 23).

In order to cover the fast-moving and growing field of AGI safety & value alignment for you, along with the new set of reviews done using LLM research tools that we publish here, we will publish surveys and reviews, or condensed versions of them. We will soon publish an entire issue of surveys and reviews.

Review: Safety at Scale: Comprehensive Survey of Large Model Safety:

> *[LLMs] are now foundational to a wide range of applications, including conversational AI, recommendation systems, autonomous driving, content generation, medical diagnostics, and scientific discovery. However, their widespread deployment also exposes them to significant safety risks, raising concerns about robustness, reliability, and ethical implications. This survey provides a systematic review of current safety research on large models, covering Vision Foundation Models (VFMs), Large Language Models (LLMs), Vision-Language Pre-training (VLP) models, Vision-Language Models (VLMs), Diffusion Models (DMs), and large-model-based Agents.*

Review: Large language Model-Powered AI Systems Achieve Self-Replication with No Human Intervention. One of the *basic AI drives* postulated by Omohundro in his classic papers was self-replication.[4] The basic AI drives – e.g. self-preservation, acquisition of resources, replication, rationality (as defined in microeconomics), preservation of utility function, and, the biggie – self-improvement – emerge from goal-seeking as they all are implied by effective goal-seeking. Bostrom

---

[2] Carroll, Sean, *The Many Hidden Worlds of Quantum Mechanics*. 2023. The Great Courses. Audio.
[3] Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality* (First edition. ed.). New York: Alfred A. Knopf.
[4] Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental and Theoretical Artificial Intelligence, 26*(3), 303-315. doi:10.1080/0952813X.2014.895111.

later gave his own analysis, calling them *instrumental drives*, i.e. drives that emerge from goal-seeking as instruments to achieve goals.[5]

Review: Large Language Models Pass the Turing Test. Besides the emergence of basic or instrumental AI drives, another metric for progress toward AGI is the Turing Test. Just one implication for safety is an AGI's ability to persuade humans is enhanced by its ability to simulate a human. The article contains the most thorough set of references covering all the important and classic papers in the history of the Turing Test that I have seen.

Review: On Regulating Downstream AI Developers. Obviously, foundation models, named for their broad range of potential downstream applications, must be regulated. Not so obvious is how to regulate the developers who adapt foundation models to their specific purpose.

Review: AI Governance through Markets. Along with Dean Ball's article, Tomei et al., lay out a menu of market mechanisms for AI regulation:

> *While current governance approaches predominantly focus on regulation, we contend that market-based mechanisms offer effective incentives for responsible AI development. We examine four emerging vectors of market governance: insurance, auditing, procurement, and due diligence, demonstrating how these mechanisms can affirm the relationship between AI risk and financial risk while addressing capital allocation inefficiencies.*

Our review also critiques their proposals.

Review: Addressing the challenges of harmonizing law and artificial intelligence technology in modern society. Posner's thesis is that the common law (e.g. contracts, torts, property) and to a degree, criminal law, evolve economically effective regulatory solutions over time.[6] That's a wonderful thesis, but since the legal system is not privatized, it moves glacially slowly. Is it even possible for the slowly-evolving legal system to respond quickly enough to the rapidly-moving AI arms race to obviate the existential risk it poses? How will new law covering essentially one or more new species of intelligence and agency interface with existing law?

---

[5] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford, England: Oxford University Press.
[6] Posner, R. A., & Aspen Publishers. (2014). *Economic analysis of law* (Ninth edition. ed.). New York: Wolters Kluwer Law & Business. Posner, Richard A. "A Theory of Negligence." Journal of Legal Studies, vol. 1, no. 1, 1972, pp. 29–96. Posner, Richard A. "The Economic Approach to Law." Texas Law Review, vol. 53, 1975, pp. 757–782.