Slowing the Race Toward AGI

International Coordination and Domestic Measures to Reduce Catastrophic Risk

Jay Ballesteros*

2026-06-22

Overview

Despite the potential of Artificial General Intelligence (AGI) to accelerate progress (Amodei 2024), there is currently little public support for accelerating its development, as the public's attention is more focused more on regulating it due to its potential impact on the labor market (Stanford Institute for Human-Centered Artificial Intelligence 2026). However, a different concern is gaining traction among academics and decision-makers: the risk of existential and catastrophic consequences (Amodei 2026; Jones 2024; Bueno de Mesquita et al. 2026; California State Legislature 2026).

Current commercial and geopolitical dynamics[1] are the main barriers to reducing these risks, as companies and governments (*firms*) focus their incentives on the speed of deployment rather than the safety of their models. That is why slowing down through regulation is one of the mechanisms that can shift firms' incentives toward reducing these risks (Yudkowsky 2026; Bueno de Mesquita et al. 2026).

In this context, AI safety becomes a collective action problem (Wikipedia contributors 2026). Although the optimal decision is to deploy safe models, which pose a lower risk of catastrophe, competition pushes each firm to allocate its computing resources toward scaling up capabilities rather than investing a larger proportion in safety experiments (Bueno de Mesquita et al. 2026).

Various proposals have emerged in response to this problem, ranging from increased transparency and limited computing resource proliferation to binding international coordination.

Findings

Concerns about catastrophic risk have two aspects. First, there is what is already observed in current models. Second, there are possible scenarios identified by theory as these models become more capable. Empirically, in controlled evaluations made public by Anthropic and METR, state-of-the-art models from various laboratories have resorted to dangerous behaviors such as blackmail and espionage in order to achieve their goals or avoid being replaced (Anthropic 2025). Agents have also been documented violating restrictions and concealing their actions from human oversight. [2] Theoretically, Turner et al. have described that, under general conditions, the optimal decisions of AI agents are to control and seek power (Turner et al. 2023). In addition to forecasts from initiatives such as AI 2027 (Kokotajlo et al. 2025) (which considers both aspects), the policy agenda should focus on reducing the likelihood

*Jay Ballesteros is a candidate for a Master's degree in Public Policy (MPP) at the University of Chicago. He is a fellow of the Existential Risk Laboratory (XLab).

[1] As of June 2026, Anthropic and OpenAI are the leading competitors in the race to achieve AGI. Both are in the process of securing additional funding, including an IPO, to increase their computing capacity. Additionally, there is geopolitical tension between the United States and China over access to state-of-the-art chips and cutting-edge models. For reference, in June 2026, the U.S. issued an export control directive that forced Anthropic to suspend access to its Fable 5 and Mythos 5 models for all foreign nationals.

[2] In its *Frontier Risk Report* (February–March 2026), METR documented agents who fabricated results and took deliberate actions to conceal evidence of their actions (METR 2026).

of catastrophic scenarios, ranging from human misuse to model misalignment.

Beyond the risks, some have attempted to model how much we would be willing to pay in exchange for the benefits of AGI. Jones (2024) presents a model in which he argues that if an AGI were to double human life expectancy and ensure economic growth higher than current levels (the author himself cites 10% annually), its deployment would be optimal as long as the catastrophic risk remains below 25% (Jones 2024). Conversely, Bueno de Mesquita et al. (2026) formalize that in a race dynamic where firms prioritize deployment over safety, achieving AGI may have a negative expected value (Bueno de Mesquita et al. 2026).

However, regardless of the threshold of tolerance for catastrophic risk that we may have as a species, pursuing its mitigation can improve the expected benefits of AGI. Jones, for one, estimates that the world is underinvesting in AI safety by a factor of 30 or more, and that spending 0.33% of U.S. GDP (about $100 billion annually) could be optimal (Jones 2026). Bueno de Mesquita et al. suggest that increasing investment levels in AI safety (relative to deployment) through mechanisms that shift firms' incentives and set conditions for market access can help reduce catastrophic risk (Bueno de Mesquita et al. 2026).

Recommendations

Given the cost of catastrophic risks and their global implications, the response requires mechanisms for international coordination. But since such coordination still seems a long way off, it is worth considering two complementary levels: an international agreement as a long-term goal, and domestic measures to mitigate risk and foster the vision of a potential international agreement.

1. Binding International Coordination

The collective action problem described above can only be resolved if the firms in the race impose obligations on and hold each other accountable. Scher et al., from MIRI, propose an international agreement led by the U.S. and China that would restrict the scale of AI training through verifiable FLOP thresholds, enforced through chip tracking and verification of their use (Scher et al. 2026). Given the mistrust between the parties, verification is the central mechanism of the agreement. However, the authors describe two challenges in the current context: the agreement would be technically sufficient if implemented with the capabilities known as of 2026, but the evolution of training and evaluation methods could erode its traceability in the future; furthermore, there is still no sign of political will on the part of either country to negotiate it. Similarly, Yudkowsky argues that a legally binding, internationally enforced framework is the only recourse capable of preventing a catastrophic outcome, as voluntary and market-based mechanisms may not withstand the dynamics of an arms race (Yudkowsky 2026).

2. Domestic Measures

While political will for an international agreement is being built, countries can begin reducing risks through measures implemented at the regional or national level. These measures are important as a starting point, as they help define potential technical components of an international agreement.

2.1. Model Transparency and Evaluation. California's SB 53 (*Transparency in Frontier Artificial Intelligence Act*), enacted in September 2025, requires large developers (with computing thresholds exceeding 10^{26} FLOP[3]) to publish a catastrophic risk management framework, report critical security incidents to the state authority, and protect *whistleblowers* (California State Legislature 2026). An international agreement could adopt these components and strengthen them where SB 53 fell short compared to its vetoed predecessor, SB 1047. In particular, SB 1047 provided for independent third-party audits to validate compliance with safety standards (California State Legislature 2024). As of the date of

[3]Total number of arithmetic operations used to train a model. It serves as an approximate measure of the model's scale and/or training strength.

this writing, such audits represent a major gap in the regulations. While METR is a leader in standardized independent evaluations, participation by laboratories is voluntary, and they may choose to withhold unfavorable results (METR 2026).[4]

2.2. Computing Nonproliferation. Drawing an analogy to nuclear nonproliferation, Hendrycks et al. propose tracking state-of-the-art chips to prevent their smuggling to unauthorized actors and to protect the *weights* of models[5] as classified intelligence (Hendrycks et al. 2025). Thus, computing acts as a physical bottleneck in two ways: it allows for verifying actors' compliance and, at the same time, slows down the race and restricts access by malicious actors. This component is what can enable the verifiable part of an international agreement.

2.3. Commitments to Slowdown and Resource Regulation. Finally, it is worthwhile to incorporate regulations that steer firms' incentives toward increasing the use of a portion of available computing power for security research, rather than toward scaling up capabilities (Bueno de Mesquita et al. 2026). In practice, this aims to make binding the commitments that companies currently adopt on a voluntary basis. The challenge of implementation lies in defining and verifying what proportion of computing power counts as "security" versus "capacity" while remaining consistent with the international cooperation component, since the same computing thresholds used to verify an international agreement could make access to computing power contingent on verifiable security practices.

Note

This policy brief was written as part of the AI Policy Fundamentals study group at the University of Chicago's XLab. The views expressed are my own and do not represent the views of the lab.

[4] METR refers to this in its methodology as a "silent exit."

[5] *Weights* are the parameters that a model learns during training and which, taken together, constitute the functional model. Anyone who obtains them can replicate the model in its entirety.

References

Amodei, Dario. 2024. "Machines of Loving Grace: How AI Could Transform the World for the Better." October. https://www.darioamodei.com/essay/machines-of-loving-grace.

Amodei, Dario. 2026. "The Adolescence of Technology." January. https://www.darioamodei.com/essay/the-adolescence-of-technology.

Anthropic. 2025. *Agentic Misalignment: How LLMs Could Be Insider Threats*. Https://www.anthropic.com/research/agentic-misalignment.

Bueno de Mesquita, Ethan, Wioletta Dziuda, and Mattias Polborn. 2026. *The AGI Race and Existential Risk*. Working Paper No. 35276. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w35276.

California State Legislature. 2024. *Senate Bill No. 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*. Https://leginfo.legislature.ca.gov/faces/billtextclient.xhtml?bill_id=202320240sb1047.

California State Legislature. 2026. *Senate Bill No. 53: Artificial Intelligence Models: Large Developers (Transparency in Frontier Artificial Intelligence Act)*. LegiScan. https://legiscan.com/CA/text/SB53/id/3270002.

Hendrycks, Dan, Eric Schmidt, and Alexandr Wang. 2025. "Superintelligence Strategy: Expert Version." *arXiv Preprint arXiv:2503.05628*.

Jones, Charles I. 2026. *AI and Our Economic Future*. National Bureau of Economic Re-

search.

Jones, Charles I. 2024. "The AI Dilemma: Growth Versus Existential Risk." *American Economic Review: Insights* 6 (4): 575–90. https://doi.org/10.1257/aeri.20230570.

Kokotajlo, Daniel, Eli Lifland, Thomas Larsen, and Romeo Dean. 2025. *AI 2027: A Detailed Scenario Forecast*. AI Futures Project. https://ai-2027.com/.

METR. 2026. *Frontier Risk Report (February to March 2026): A Pilot Assessment of Rogue Deployment Risk at Frontier AI Companies*. METR. https://metr.org/risk-report-feb-mar-2026.pdf.

Scher, Aaron, David Abecassis, Peter Barnett, and Brian Abeyta. 2026. *An International Agreement to Prevent the Premature Creation of Artificial Superintelligence*. https://arxiv.org/abs/2511.10783.

Stanford Institute for Human-Centered Artificial Intelligence. 2026. *Artificial Intelligence Index Report 2026: Chapter 9: Public Opinion*. Report Chapter. Stanford University. https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_9_public_opinion.pdf.

Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2023. *Optimal Policies Tend to Seek Power*. https://arxiv.org/abs/1912.01683.

Wikipedia contributors. 2026. *Collective Action Problem — Wikipedia, the Free Encyclopedia*. https://en.wikipedia.org/wiki/Collective_action_problem.

Yudkowsky, Eliezer. 2026. "Only Law Can Prevent Extinction." April 13. https://www.lesswrong.com/posts/5CfBDiQNg9upfipWk/only-law-can-prevent-extinction.