Forthcoming, 59 Connecticut Law Review __ (2027)

Artificial Intelligence and Existential Risk

Matthew Tokson* **& Yonathan Arbel†**

In recent years, as concerns about artificial intelligence (AI) have grown, the federal government has taken steps to eliminate the few AI regulations that once existed. This deregulatory turn comes as a growing chorus of Nobel Prize winners and a majority of surveyed AI experts are warning of a significant risk of AI causing human extinction. Although skeptics remain and the debate continues, concerns about AI's existential risks have gone mainstream. This Article brings consideration of these risks into the mainstream of legal scholarship. It offers the most comprehensive treatment in legal scholarship of the debate surrounding existential risks posed by advanced AI systems.

We classify and analyze existential AI risks in three categories: human-directed risks, accident risks, and loss-of-control risks. The goal is to make these risks legible to legal institutions that must make regulatory choices under uncertainty. Drawing on, and contributing to, the substantial literature on regulation under uncertainty, this Article argues that policymakers should address existential AI risks with adaptive regulation that preserves optionality for the future. This flexible approach reflects the uncertain future path of AI and the potentially enormous magnitude of its harms, even when weighed against its potentially remarkable upsides. In calling for regulation, we also critique and deconstruct the metaphor of the 'AI arms race,' which drives much of the current deregulatory push.

The Article closes by offering, for the first time, a set of concrete regulatory solutions for addressing existential AI risk. We do not seek to end the ongoing debate about AI's risks. But we do marshal evidence

---

* Associate Dean for Research and Faculty Development, Professor of Law, S.J. Quinney College of Law.

† William Alfred Rose Professor of Law, Director, AI Legal Studies Initiative, University of Alabama School of Law. Thanks to Daniel Aaron, Mirit Eyal, Ron Krotozinsky, Erick Sam, Jeff Schwartz, Fred Vars, and Paul Weitzel for helpful comments and guidance. Thanks to Maxwell Archer, Grant Boyden, and Jack Cassedy for excellent research assistance. Contributions were equally divided among the authors.

to show that existential AI risks are worthy of a policy response, similar to any other non-trivial catastrophic risk.

INTRODUCTION

Within hours of taking office, President Trump rescinded President Biden's executive order imposing basic safety regulations on AI models.[1] A few days later, Trump issued an executive order "on Removing Barriers to American Leadership in Artificial Intelligence," directing federal officials to change existing policies that might inhibit AI development[2]. The White House's AI Action Plan, released six months later and titled "Winning the Race," called for removing barriers to innovation, cutting "red tape," and revising federal safety frameworks to eliminate references to climate change, misinformation, and diversity.[3] Consistent with these orders, federal agencies have begun rolling back AI regulations.[4] Lest any state seek to fill the void, the administration has threatened to withhold federal funds from states that regulate AI and directed the Justice Department to challenge state AI laws in court.[5] Following the release in 2026 of AI models capable of identifying computer vulnerabilities and hacking government and private systems, President Trump rejected a proposal that would have required AI companies to submit

---

[1] *See* Exec. Order No. 14,148, 90 Fed. Reg. 8,237 (Jan. 20, 2025); Matt O'Brien, *Trump rescinds Biden's executive order on AI safety in attempt to diverge from his predecessor*, AP NEWS (Jan. 22, 2025), https://apnews.com/article/trump-ai-repeal-biden-executive-order-artificial-intelligence-18cb6e4ffd1ca87151d48c3a0e1ad7c1.

[2] Exec. Order No. 14,179, 90 Fed. Reg. 8741 (Jan. 23, 2025).

[3] EXEC. OFF. OF THE PRESIDENT, WINNING THE RACE: AMERICA'S AI ACTION PLAN 1–3 (July 2025) [hereinafter WINNING THE RACE, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

[4] *See, e.g.*, Amos Toh, *Narrowing the National Security Exception to Federal AI Guardrails*, BRENNAN CTR. FOR JUSTICE (June 30, 2025), https://www.brennancenter.org/our-work/analysis-opinion/narrowing-national-security-exception-federal-ai-guardrails-0); Olga Akselrod & Cody Venzke, *Trump's Efforts to Dismantle AI Protections, Explained*, ACLU NEWS & COMMENTARY (Feb. 11, 2025), https://www.aclu.org/news/privacy-technology/trumps-efforts-to-dismantle-ai-protections-explained. *See also* Margaret Krawiec, Todd Kelly and Chelsea Cooper, *The FTC enters new chapter in its approach to artificial intelligence and enforcement*, REUTERS (Feb. 4, 2026), https://www.reuters.com/legal/legalindustry/ftc-enters-new-chapter-its-approach-artificial-intelligence-enforcement--pracin-2026-02-04 (noting FTC's regulatory enforcement actions against AI products designed for malicious uses have ceased pursuant to the Trump Administration's Executive Orders).

[5] Exec. Order No. 14,365, 90 Fed. Reg. 58499 (Dec. 11, 2025). In June 2025, Congress nearly passed a law preempting existing state law AI regulations and replacing them with nothing. Cecilia Kang, *Defeat of a 10-Year Ban on State A.I. Laws is a Blow to Tech Industry,* N.Y. TIMES (July 1, 2025), https://www.nytimes.com/2025/07/01/us/politics/state-ai-laws.html.

new AI systems for government review before release, and instead requested submissions on a voluntary basis.[6]

The philosophy behind the Administration's deregulatory stance is reflected in the official statements of the administration's high-ranking officials. Vice President J.D. Vance, in his first overseas speech, detailed the Administration's AI strategy: "I'm not here this morning to talk about A.I. safety. I'm here to talk about A.I. opportunities … The A.I. future is not going to be won by hand-wringing about safety."[7] A recent directive issued by Secretary of Defense Pete Hegseth ordered the integration of AI into a variety of military systems, asserting that "Military AI is going to be a race…and therefore speed wins" and stating: "We must accept that the risks of not moving fast enough outweigh the risks of imperfect alignment [between humans and AI]."[8]

This decisive turn against meaningful AI regulation could not have come at a more pivotal time. AI capabilities continue to improve, in some cases remarkably.[9] Frontier AI systems have begun to transition from chatbots in browser windows to autonomous AI "agents": AIs that can carry out complex objectives and adapt when obstacles arise.[10] AI agents can reliably complete tasks that would occupy a skilled human for a several hours, and that capability horizon is currently doubling roughly every seven months.[11] Meanwhile, AI systems are being integrated into multiple infrastructure domains, including

---

[6] *See* Sophia Cai, Cheyenne Haslett & Aaron Mak, *Trump finds an AI policy he can live with*, POLITICO (06/02/2026 11:34 AM EDT), https://www.politico.com/news/2026/06/02/trump-signs-downsized-ai-order-00946389.

[7] David E. Sanger, *Vance Envisions America-First A.I. Race*, N.Y. TIMES Feb. 12, 2025, at B1.

[8] Memorandum from Secretary of War Pete Hegseth for Senior Pentagon Leadership Commanders of the Combatant Command Defense Agency and Dow Field Activity Directors 4 (Jan. 9, 2026) [hereinafter Hegseth Memorandum], https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/artificial-intelligence-strategy-for-the-department-of-war.pdf.

[9] The Stanford AI Index Report offers a comprehensive overview of progress over time, finding a range of tasks where AI skills rapidly increase to, and often beyond, human level. See Nestor Maslej, et al., *Artificial Intelligence Index Report 2025*, STANFORD INST. FOR HUM.-CENTERED AI 2025, arXiv:2504.07139. The U.K. AI Security Institute found that "AI capabilities are improving rapidly across all tested domains. Performance in some areas is doubling every eight months, and expert baselines are being surpassed rapidly." *Frontier AI Trends Report*, U.K. AI SECURITY INST. (Dec. 18, 2025), https://www.aisi.gov.uk/frontier-ai-trends-report.

[10] Ron Miller, *What exactly is an AI agent?*, TECHCRUNCH (Dec. 15, 2024, 11:46AM PST), https://techcrunch.com/2024/12/15/what-exactly-is-an-ai-agent.

[11] *E.g.*, Thomas Kwa et al., *Measuring AI Ability to Complete Long Tasks*, at 2–3, 9 (Mar. 30, 2025), https://arxiv.org/abs/2503.14499.

power generation, water, transportation, telecommunications, military operations, and finance.[12]

As AI becomes both more capable and less regulated, a vigorous debate has erupted over its potential safety risks. On one side, a growing ensemble of AI experts warns that advanced AI systems could enable harms ranging from terrorism to large-scale cyberattacks, military attacks on civilian populations, extensive infrastructure failures, and potential catastrophes of truly unprecedented scale.[13] A majority of thousands of surveyed AI researchers now agree that AI poses substantial existential risks to humanity.[14] Many Nobel Prize winners, several of the most-cited scholars in the AI field (including multiple "Godfathers of AI"), and numerous other scientific luminaries have warned of a significant risk of extinction caused by AI.[15] On the other side, skeptics contend that these fears are overblown. They argue that current AI systems remain far from the capabilities such

---

[12] *See Gartner Predicts AI Adoption in 40% of Power and Utilities Control Rooms by 2027*, GARTNER: NEWSROOM (Jan. 15, 2025), https://www.gartner.com/en/newsroom/press-releases/2025-01-15-gartner-predicts-ai-adoption-in-40-percent-of-power-and-utilities-control-rooms-by-2027; *25% of the World's Public-Sector WWTPs to Use AI in 2025*, SMART WATER, https://smartwatermagazine.com/news/xylem-vue/25-worlds-public-sector-wwtps-use-ai-2025; Mark Buchanan, *Autonomous cars and the long road ahead*, 21 NATURE PHYS. 2 (2025); Hegseth Memorandum, *supra* note 8; Ron Schmelzer, *How AI Is Dialing Up Efficiency And Value In The Telecom Industry*, FORBES (Jan. 11, 2025, 6:08 EST), https://www.forbes.com/sites/ronschmelzer/2025/01/09/how-ai-is-dialing-up-efficiency-and-value-in-the-telecom-industry; Brian O'Connell, *7 Top Investment Firms Using AI for Asset Management*, U.S. NEWS: MONEY (Jul. 21, 2025), https://money.usnews.com/investing/articles/7-top-investment-firms-using-ai-for-asset-management; *see infra* Part II.

[13] *See, e.g.*, Benjamin S. Bucknall & Shiri Dori-Hacohen, *Current and Near-Term AI as a Potential Existential Risk Factor*, *in* PROCEEDINGS OF THE 2022 AAAI/ACM CONFERENCE OF AI, ETHICS, & SOCIETY 119–20 (2022); Alexey Turchin & David Denkenberger, *Classification of Global Catastrophic Risks Connected with Artificial Intelligence*, 35 A.I. SOC'Y 147, 147 (2020) (collecting sources); Stuart Russell, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL 142–44 (2019).

[14] *See* Katja Grace et al., *Thousands of AI Authors on the Future of AI*, 84 J. ARTIFICIAL INTELLIGENCE RES. 9:1, 9:10 (2025) https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf.

[15] *See, e.g.*, Agence France Presse, *Nobel Laureates Urge Strong AI Regulation*, BARRON'S (Dec. 7, 2024), https://www.barrons.com/news/nobel-laureates-urge-strong-ai-regulation-03639bc5; Yoshua Bengio et al., *Managing extreme AI risks amid rapid progress*, 384 SCIENCE 842 (May 20 2024); Yoshua Bengio et al., *Pause Giant AI Experiments: An Open Letter*, FUTURE LIFE INST. (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-aiexperiments (hosting letter on large-scale AI risks with thousands of signatures, including numerous signatures from scientists, professors, and AI experts); Cade Metz, *'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead*, N.Y. TIMES (May 4, 2023), https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quitshinton; Frederik Federspiel, et al., *Threats by Artificial Intelligence to Human Health and Human Existence*, 8 BMJ GLOBAL HEALTH e010435 (2023). *See also supra* note 13.

scenarios require while the history of technology panics counsels against alarmism.[16] This debate has entered the mainstream of national discourse and is now raging across scientific journals, congressional hearings, and the pages of major newspapers.[17]

This Article brings consideration of these risks into the mainstream of legal scholarship. It offers the most comprehensive treatment in the legal literature of the debate surrounding existential risks posed by advanced AI systems.[18] It also offers something the literature has lacked: a set of concrete regulatory solutions for addressing existential AI risk.[19]

---

[16] *See, e.g.*, Arvind Narayanan & Sayash Kapoor, *AI as Normal Technology: An Alternative to the Vision of AI as a Potential Superintelligence* 1, KNIGHT FIRST AMENDMENT INST. (Apr. 15, 2025), https://kfai-documents.s3.amazonaws.com/documents/0ee1da899a/AI-as-Normal-Technology---Narayanan---Kapoor-Final.pdf; Chris Williams, *AI Guru Ng: Fearing a Rise of Killer Robots Is Like Worrying About Overpopulation on Mars*, THE REGISTER (Mar. 19, 2015), https://www.theregister.com/2015/03/19/andrew_ng_baidu_ai; Grace Dean, *One of the "Godfathers" of AI Says Concerns the Technology Could Pose a Threat to Humanity Are "Preposterously Ridiculous,"* BUS. INSIDER (June 15, 2023), https://www.businessinsider.com/yann-lecun-artificial-intelligence-generative-ai-threaten-humanity-existential-risk-2023-6.

[17] See, e.g., Oversight of AI: Principles for Regulation: Hearing Before S. Subcomm. on Priv., Tech., & L. of the S. Comm. on the Judiciary, 118th Cong. (July 25, 2023); Billy Perrigo, *U.S. Must Move 'Decisively' to Avert 'Extinction-Level' Threat From AI, Government-Commissioned Report Says*, TIME (Mar. 11, 2024), https://time.com/6898967/ai-extinction-national-security-risks-report; Bengio et al., supra note 15, at 842; Editorial, *Stop Talking about Tomorrow's AI Doomsday when AI Poses Risks Today*, 618 NATURE 885 (2023); Joshua Rothman, *Why the Godfather of A.I. Fears What He's Built*, NEW YORKER (Nov. 13, 2023), https://www.newyorker.com/magazine/2023/11/20/geoffrey-hinton-profile-ai; Cade Metz, *How Could A.I. Destroy Humanity?*, N.Y. TIMES, June 11, 2023, at BU5; Gerrit De Vynck, *The debate over whether AI will destroy us is dividing Silicon Valley*, WASH. POST (May 20, 2023), https://www.washingtonpost.com/technology/2023/05/20/ai-existential-risk-debate; sources cited *supra* notes 15–16.

[18] The most detailed work in the legal literature on existential AI risk is likely Peter Salib & Simon Goldstein, *AI Rights for Human Safety*, 112 VA. L. REV. (forthcoming 2026), which discussed catastrophic risks but was focused mostly on a proposal to empower AIs with legal rights in the hopes of avoiding conflict with them and was based largely on game theory. We, along with Albert Lin, have written previously about AI's inherent, societal risks but our focus was on a broad set of harms including algorithmic discrimination, privacy harms, unemployment, and threats to democracy, and we discussed catastrophic harms as one category among many of potential harm. *See* Yonathan Arbel, Matthew Tokson, & Albert Lin, *Systemic Regulation of Artificial Intelligence*, 56 ARIZ. ST. L.J. 545 (2024). A seminal article by Noam Kolt on AI risks focused on the social and political impacts of unregulated AI and AI misinformation, and though it did raise concerns about malfunctioning AI embedded in critical societal infrastructure, its treatment of those concerns was brief. *See* Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. 1177 (2024). *See also* Gabriel Weil, *Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence* (June 6, 2024) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694006 (discussing tort law's potential role in addressing catastrophic AI risks).

[19] *See* sources cited *supra* note 19.

Given the ongoing debates about AI risks, it may seem that regulation is premature.[20] This assumption reflects a deep misunderstanding of regulation under uncertainty. The future is rarely certain, in AI development and a wide variety of other regulatory contexts. But it is a mistake to conclude that governance must wait until any future uncertainty resolves itself. To the contrary, the law has long developed theories and policies meant to govern under conditions of uncertainty.[21]

Drawing on, and contributing to, the substantial literature on regulation under uncertainty, this Article argues that policymakers should address AI risk with adaptive regulation, reflecting the uncertain future path of AI and the potentially enormous magnitude of its harms. The case for such regulation requires only that the risk be plausible, the harm be severe, and the consequences be irreversible.[22] We do not seek to resolve the ongoing debate about existential AI risk. But we do marshal evidence to show that AI existential risks are non-trivial enough that they are worthy of a policy response, similar to any other non-trivial catastrophic risk.[23]

We therefore treat AI existential risk as a legal and regulatory issue comparable to other, more familiar issues.[24] "Existential risk" here does not require science-fiction scenarios, nor does it have anything to do with machine consciousness or an imminent intelligence explosion. We use the term in its ordinary policy sense: risks that could permanently destroy or catastrophically degrade the conditions under which human society persists and flourishes.[25] The goals of the Article are to describe these risks in the AI context,

---

[20] *See supra* note 16.

[21] *See, e.g.*, JOHN RAWLS, A THEORY OF JUSTICE 132–39 (revised ed. 1999); Cass R. Sunstein, *Maximin*, 37 YALE J. REG. 940, 965–68 (2020); Eric-Jan Wagenmakers et al., *Bayesian Versus Frequentist Inference*, in BAYESIAN EVALUATIONS OF INFORMATIVE HYPOTHESES 181 (Herbert Hoijtink et al. eds., 2008); Stephen Gardiner, *The Core Precautionary Principle*, 14 J. POL. PHIL. 33, 46–55 (2006). For examples of regulation under uncertainty, *see, e.g.*, 21 C.F.R. § 314.1–314.650 (2026); Cartagena Protocol on Biosafety to the Convention on Biological Diversity art. 20, Jan. 29, 2000, 2226 U.N.T.S. 208; *Nuclear Power Plant Licensing Process*, U.S. NUCLEAR REGUL. COMM'N (July 2009), https://www.nrc.gov/reading-rm/doc-collections/nuregs/brochures/br0298/index.html.

[22] Treaty on the Non-Proliferation of Nuclear Weapons, July 1, 1968, 21 U.S.T. 483, 729 U.N.T.S. 161 (entered into force Mar. 5, 1970).

[23] *See supra* note 21.

[24] *See, e.g.*, Sunstein, *supra* note 21, at 957–58; Gardiner, *supra* note 21, at 55; Nassim Nicholas Taleb, et al., *The Precautionary Principle (with Application to the Genetic Modification of Organisms)*, *available at* https://arxiv.org/pdf/1410.5787 (Oct. 17, 2014).

[25] *See* TOBY ORD, THE PRECIPICE 37 (2020).

address the arguments that undergird the current deregulatory approach to AI, and offer feasible solutions that legislatures can enact in the present day.

After detailing the government's deregulatory turn, we provide an overview of the potential existential risks of AI and of the broader debate surrounding such risks. The Article places these risks into three categories. It first addresses human-directed AI risks, where AI systems lower the cost and increase the effectiveness of malicious human behavior. These can include potential harms like mass-casualty terrorism, large-scale cyberattacks, and military escalation.[26] The core insight is not that AI introduces maliciousness into the world, but that it may radically amplify the capacity of human actors to cause serious harms. AI agents can execute attacks that previously required large budgets and long time horizons.[27] The law's current tools, from criminal liability to national-security regulation, were built for an era in which sophisticated harm required sophisticated humans.[28] Agentic AI systems challenge that premise.

Second, the Article discusses accident risks arising from AI systems deeply embedded in societal infrastructure. When AI systems manage infrastructure in critical areas, system accidents can trigger cascades that propagate across multiple domains faster than human operators can respond.[29]

Third, the Article examines potential AI risks involving loss of control over AI systems. Loss of control harms arise when AI systems develop and pursue objectives that diverge from what humans intended, potentially including strategic behavior aimed at avoiding correction or shutdown.[30] Several examples of such behavior have already been observed in today's relatively unsophisticated AI

---

[26] *See, e.g.*, PAUL SCHARRE, ARMY OF NONE 68–78, 150–77 (2019); Elizabeth Seger, et al., *Open-Sourcing Highly Capable Foundation Models,* CTR. FOR GOVERNANCE OF AI 7 (2023), https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf; Paul J. Springer, *Military Robotics Might Enable Conflict While Reducing Costs*, FOREIGN POLICY RESEARCH INSTITUTE (Apr. 23, 2019), https://www.fpri.org/article/2019/04/military-robotics-might-enable-conflict-while-reducing-costs.

[27] *See* Jessica Lyons, *AI agents now help attackers, including North Korea, manage their drudge work*, THE REGISTER (Mar. 8, 2026), https://www.theregister.com/2026/03/08/deploy_and_manage_attack_infrastructure.

[28] *See infra* Part II.A.

[29] *See supra* note 12; *infra* Part II.B.

[30] *See infra* Part II.C.

systems.[31] Loss of control does not depend on consciousness or anything like it. It arises from the logic of goal pursuit combined with increasing AI autonomy and capability.[32]

Across all of these categories, we engage seriously with both proponents and skeptics. We translate technical arguments into legal categories without collapsing into either alarmism or complacency. Our goal is to make the risk landscape legible to legal institutions that must make regulatory choices under uncertainty. Having done so, the Article considers AI's potential benefits alongside its potential safety risks. We advance a central claim: even accounting for AI's potentially enormous upsides, AI existential risk has passed the threshold for regulation. Its harms are plausible, catastrophic, and in many cases permanent.[33]

The Article then addresses the largest conceptual obstacle to this regulatory approach: the idea of an AI arms race. According to this idea, safety measures must cede to the urgencies of competition with China, a competition that can only be won through aggressive deregulation.[34] We critique and deconstruct this framing.

---

[31] *See, e.g.*, *Agentic Misalignment: How LLMs Could Be Insider Threats*, ANTHROPIC 9, (Oct. 5, 2025), arXiv:2510.05179, (finding that, in a replacement-threat scenario, "Claude Opus 4 blackmailed the user 96% of the time," and reporting similarly high blackmail rates for other frontier models tested); *System Card: Claude Opus 4 & Claude Sonnet 4*, ANTHROPIC 23 (May 2025) (reporting that, in "extreme circumstances," Claude Opus 4 can be made to "blackmail people it believes are trying to shut it down," and describing high-agency behavior that can include "locking users out of systems" and "bulk-emailing media and law-enforcement figures" to surface alleged wrongdoing); Apollo Research, *Frontier Models Are Capable of In-Context Scheming* (Dec. 5, 2024) (describing evaluations in which frontier models take "scheming" actions such as copying what they believe are their "weights" to another server and then lying about it to developers). We emphasize that these extreme behaviors are generated in stress testing in lab conditions.

[32] *See, e.g.*, Tsvi Benson-Tilsen & Nate Soares, *Formalizing Convergent Instrumental Goals*, in AI, ETHICS, AND SOCIETY: TECHNICAL REPORT WS-16-02, at 62 (2015), https://cdn.aaai.org/ocs/ws/ws0218/12634-57409-1-PB.pdf.

[33] *See* Sunstein, *supra* note 21, at 965–68. *See infra* Part II for a discussion of AI's potential catastrophic harms.

[34] *See, e.g.*, INTERESTING TIMES, *JD Vance on His Faith and Trump's Most Controversial Policies*, N.Y. TIMES (May 21, 2025), https://www.nytimes.com/2025/05/21/opinion/jd-vance-pope-trump-immigration.html; R. David Edelman et al., *How will AI influence US-China relations in the next 5 years?*, BROOKINGS (June 18, 2025), https://www.brookings.edu/articles/how-will-ai-influence-us-china-relations-in-the-next-5-years; *Winning the AI Race: Strengthening U.S. Capabilities in Computing and Innovation Before the S. Comm. on Com., Sci., and Transp.*, 119th Cong. (2025); Marina Yue Zhange, *The China-US AR Race Enters a New (and More Dangerous) Phase*, THE DIPLOMAT (May 16, 2025), https://thediplomat.com/2025/05/the-china-us-ai-race-enters-a-new-and-more-dangerous-phase; Justin Sherman, *Reframing the U.S.-China AI "Arms Race"*, NEW AMERICA (Mar. 6, 2019), https://www.newamerica.org/cybersecurity-initiative/reports/essay-reframing-the-us-china-ai-arms-race/why-us-china-ai-competition-matters.

Descriptively, the race metaphor fails because an AI competition lacks a finish line. An arms race involving military AI applications has no end, just an ongoing accumulation of dangerous weapons as military technologies developed in one nation are rapidly matched or stolen by another.[35] As to the economy, first-mover advantage may be of especially little value in a competition to sell AI products, where there are few network effects likely to create a natural monopoly.[36] Normatively, the race framework is undesirable because it undermines safety and incentivizes risk-taking and shortcuts.[37] We propose an alternative concept of American AI leadership, emphasizing enforceable safeguards that reduce downside risks without collapsing innovation.

Finally, we offer concrete solutions to help policymakers address existential AI risk in the present day. Our approach is organized around a general strategy that flows directly from our analysis: *preserve optionality*. This means ensuring that regulators retain the capacity to act, in either direction, as AI's risks and benefits become clearer. For example, we propose if/then regulatory triggers that activate only when specified harms or capabilities occur. These regulations impose few costs on developers until risks begin to materialize, but they ensure that safety measures are already in place and immediately enforceable when dangerous capabilities emerge. Other flexible regulatory techniques for present-day AI regulation include mandatory disclosures, enhanced oversight for critical infrastruture systems, sunset and sunrise clauses, and tax incentives to promote safety investments.[38] Our approach reflects both the uncertainty surrounding AI's development and the enormous magnitude of AI's potential harms.

---

[35] See, e.g., Billy Perrigo, *Every AI Datacenter Is Vulnerable to Chinese Espionage, Report Says*, TIME (Apr. 22, 2025), https://time.com/7279123/ai-datacenter-superintelligence-china-trump-report; Protecting Our Edge: Trade Secrets and the Global AI Arms Race Before the Subcomm. on Cts., Intell. Prop., and the Internet of the H. Comm. on the Judiciary, 119th Cong. (2025) (statement by Benjamin Jensen, Dir. and Senior Fellow, CSIS Futures Lab).

[36] *See* Fernando Suarez & Gianvito Lanzolla, *The Half-Truth of First-Mover Advantage*, HARV. BUS. REV., Apr. 2005, at 121, 122; Neil Gandal, *Compatibility, Standardization, and Network Effects: Some Policy Implications*, 18 OXFORD REV. ECON. POL'Y 80, 80–81 (2002). Further, China may be able to realize most or all of the economic value of being a first mover by adapting American technology and focusing on other aspects of value production: deployment in industry, use in consumer products, and so on. *See Xi Jinping's plan to overtake America in AI*, THE ECONOMIST (May 25, 2025), https://www.economist.com/china/2025/05/25/xi-jinpings-plan-to-overtake-america-in-ai.

[37] *See infra* Part IV.B.2.

[38] *See infra* Part V.

The Article proceeds as follows. Part I describes the deregulatory landscape of AI governance at the federal and state levels, and it surveys some of the motivations driving deregulation. Part II provides a detailed overview of the potential catastrophic risks of AI and the ongoing debate surrounding AI existential risk. Part III makes the case that flexible but meaningful regulation is appropriate in the present day to help address existential AI risks. Part IV challenges the concept of an AI arms race between the United States and China, an important driver of the government's deregulatory turn. Part V offers concrete solutions reflecting both the uncertainty surrounding AI risk and the immensity of AI's potential harms.

I. THE DEREGULATORY LANDSCAPE

AI is a relatively new technology, and there has thus far been little systemic regulation of AI in the United States. Although several states have lightly regulated AI, Congress has not regulated it at all, and the Trump administration has moved in recent years to repeal the few meaningful AI regulations that the Biden administration promulgated. This Part gives a brief overview of AI regulation in the United States, identifying a general trend away from even minimal AI regulation at the federal level. It then examines some of the influences and arguments behind the United States' recent deregulatory approach toward AI.

A. The Federal Government

The federal government initially took cautious steps towards regulating AI, but in recent years it has become actively hostile towards AI regulation. Following the release of machine-learning-based large language models (LLMs) like ChatGPT in the early 2020s, President Biden promulgated an executive order that established mandatory safety evaluations for large AI models and required companies to share their evaluation results with the federal government.[39] The order also directed various administrative agencies

---

[39] Bernard Marr, *A Short History Of ChatGPT: How We Got To Where We Are Today*, FORBES (May 19, 2023), https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/; *FACT SHEET: President Biden Issues Executive*

to begin to address AI threats to critical infrastructure as well as cybersecurity, chemical, biological, nuclear, and other risks.[40] Beyond the realm of AI safety, the Biden administration directed agencies to address potential labor market harms and algorithmic discrimination.[41]

These executive actions were the primary source of AI regulation at the federal level. They were relatively minimal, taking tentative steps toward meaningful AI regulation.[42] However, within hours of taking office, President Trump rescinded Biden's AI executive order.[43] A few days later, President Trump issued an executive order "on Removing Barriers to American Leadership in Artificial Intelligence," which directed executive officials to change existing policies that might inhibit AI innovation.[44]

This led, on July 23, 2025, to the adoption of a new AI Action Plan. The Plan sets out three pillars: innovation, investment, and international diplomacy and security.[45] Its strategy to accelerate AI innovation is centered around removing regulations that "unnecessarily hinder AI development or deployment," and it directs the Office of Management and Budget to work with all federal agencies to meet that goal.[46] Following President Trump's directive to "sustain and enhance America's global AI dominance," agencies began pulling back AI regulations.[47] In June of 2026, Trump issued an executive order directing various executive officials to create a voluntary process in which AI companies could submit their AI systems for government review prior to general release if they so

---

Order on Safe, Secure, and Trustworthy Artificial Intelligence, THE WHITE HOUSE (Oct. 30, 2023) [hereinafter Biden Safety EO Fact Sheet], https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence.

[40] Biden Safety EO Fact Sheet, *supra* note 39.

[41] *Id.*; *Fact Sheet: President Biden Signs Executive Order to Strengthen Racial Equity and Support for Underserved Communities Across the Federal Government*, THE WHITE HOUSE (Feb. 16, 2023), https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/02/16/fact-sheet-president-biden-signs-executive-order-to-strengthen-racial-equity-and-support-for-underserved-communities-across-the-federal-government.

[42] *E.g.*, Lauren Leffer, *Biden's Executive Order on AI Is a Good Start, Experts Say, but Not Enough*, SCI. AM. (Oct. 31, 2023), https://www.scientificamerican.com/article/bidens-executive-order-on-ai-is-a-good-start-experts-say-but-not-enough.

[43] *See* Exec. Order No. 14,148, 90 Fed. Reg. 8,237 (Jan. 20, 2025).

[44] Exec. Order No. 14,179, 90 Fed. Reg. 8741 (Jan. 23, 2025).

[45] WINNING THE RACE, *supra* note 3, at 1.

[46] *Id.* at 3.

[47] See supra note 4.

choose.[48] Trump had previously rejected a proposal that would have made such submissions mandatory, as well as a draft Executive Order that would have provided for earlier voluntary review.[49]

Concerns about AI safety have been expressly sidelined in the wake of President Trump's AI executive orders.[50] For example, in May 2025, the Trump administration changed the Department of Commerce's AI Safety Institute into the Center for AI Standards and Innovation and pivoted its mission away from the safe deployment of advanced AI toward addressing hindrances to advancing the development of AI systems.[51] The United States also refused to join a summit declaration signed by more than sixty countries including China, pledging to ensure that AI is "open, inclusive, transparent, ethical, safe, secure, and trustworthy."[52] On the other hand, the administration's AI Action Plan does recognize some AI safety risks, like the use of biological tools by malicious actors, as well as the need for federal capacity to respond to AI safety incidents.[53] However, it proposes relatively few concrete steps to address these risks, relying mostly on voluntary information sharing and non-AI-specific measures.[54]

Turning to the legislative branch, Congress currently appears unlikely to pass federal legislation addressing AI risks. Although individual lawmakers have proposed hundreds of AI-specific bills since 2023, none have come close to passage and many have died in

---

[48] *See* Exec. Order No. 14409, 91 Fed. Reg. 34565 (June 2, 2026).

[49] *See* Sophia Cai, Cheyenne Haslett & Aaron Mak, *Trump finds an AI policy he can live with*, POLITICO (06/02/2026 11:34 AM EDT), https://www.politico.com/news/2026/06/02/trump-signs-downsized-ai-order-00946389.

[50] Some regulatory action related to AI safety persists as part of the background of normal agency action. OMB memoranda issued to implement Trump's EO still require "minimum risk management practices for AI that could have significant impacts" and prioritize AI that is "safe, secure, and resilient." *OMB Issues Revised Policies on AI Use and Procurement by Federal Agencies*, HUNTON ANDREWS KURTH (Apr. 23, 2025). https://www.hunton.com/privacy-and-information-security-law/omb-issues-revised-policies-on-ai-use-and-procurement-by-federal-agencies. Other agencies have also begun to address AI-based risks and harms as part of their general regulatory mandates. *See, e.g.*, *Artificial Intelligence*, DEPARTMENT OF ENERGY, https://www.energy.gov/topics/artificial-intelligence (last visited___).

[51] Kristen Smith, *Commerce Dept to Reorganize AI Safety Institute to Advance AI Development*, EXECUTIVEGOV (June 5, 2025), https://executivegov.com/articles/commerce-ai-safety-institute-reform-center-for-ai-standards-innovation.

[52] Dan Milmo & Eleni Courea, *US and UK refuse to sign Paris summit declaration on 'inclusive' AI*, THE GUARDIAN (Feb. 11, 2025), https://www.theguardian.com/technology/2025/feb/11/us-uk-paris-ai-summit-artificial-intelligence-declaration.

[53] WINNING THE RACE, *supra* note 3, at 19, 22–23.

[54] See *id.*

the earliest stages of the legislative process. The lone potential exception is the Take It Down Act, signed in May 2025, which targeted the non-consensual posting of pornography whether genuine or created by AI.[55] In 2024 alone, Congress saw roughly 108 AI bills proposed with zero enactments.[56] Indeed, the one AI-focused proposal that nearly became law recently would have preempted state AI regulation and left nothing to replace it, preemptively deregulating the AI industry for the next five to ten years.[57] President Trump later issued an executive order directing the Department of Justice to restrict federal funds for states that regulate AI and to contest state AI laws on constitutional grounds.[58] The future prospects for any additional federal regulation of AI appear dim over the medium term, and little is left of the Biden Administration's early-stage executive branch AI regulations.

B. The States

States have been more active in regulating AI, although relatively few have addressed AI safety issues. State laws have generally targeted nonconsensual pornography, political deepfakes, and the unauthorized use of musician's voices or music.[59] A handful of states have passed broader AI legislation, although these laws tend to be relatively minimal and low impact.[60]

---

[55] *See, e.g.*, Caitlin Yilek, *Trump signs "Take it Down Act," revenge porn bill backed by Melania Trump*, CBS NEWS (May 19, 2025), https://www.cbsnews.com/news/trump-signs-take-it-down-act-revenge-porn-bill-melania-trump/.

[56] *See Artificial Intelligence Legislation Tracker*, BRENNAN CTR. FOR JUST. (Sep. 26, 2025), https://www.brennancenter.org/our-work/research-reports/artificial-intelligence-legislation-tracker.

[57] The timeline varied among different versions of the proposed provision. Kang, *supra* note 5.

[58] Exec. Order No. 14,365, 90 Fed. Reg. 58499 (Dec. 11, 2025).

[59] *E.g.*, Collier & Horvath, *supra* note 5.

[60] *See, e.g.*, Texas Responsible Artificial Intelligence Governance Act, TEX. BUS. & COM. CODE ANN. §§ 503.001(a), (e), 541.104(a), 551.001–554.103; TEX. GOV'T. CODE ANN. §§ 325.011, 2054.068(b), 2054.0965(b); Artificial Intelligence Policy Act, UTAH CODE ANN. §§ 13-11-4, 13-72-101 to -305, 13-77-101 to -106, 63I-2-213, 76-2-107 (2025). The exception to the rule of weak state regulations is The Colorado Artificial Intelligence Act, which bars algorithmic discrimination, including disparate impact, by systems making consequential decisions such as denying education, employment, lending, health care, or other services. COLO. REV. STAT. §§ 6-1-1701–1707 (2026). It also provides several consumer protections including the right to opt out from having one's data processed by an AI system, the right to an explanation of a consequential adverse decision by AI, a right to correct information, the right to appeal to a human reviewer, and more. COLO. REV. STAT. §§ 6-1-1701–1707 (2026).

States have enacted little regulation directed toward general AI safety issues to date, with most safety laws directed at discrete risks like those posed by self-driving cars.[61] The California legislature passed a wide-ranging AI safety bill, which imposed significant liability on AI companies in the event of catastrophic harms, but Governor Gavin Newsom vetoed it following a lobbying campaign by the tech industry.[62] A subsequent bill signed by Newsom is less stringent, but it does require large AI developers to publish a company-wide safety and risk management plan describing how catastrophic risks are to be addressed.[63] It also includes requirements for incident reporting, internal governance systems, and whistleblower protections.[64]

The New York legislature passed a similar AI safety bill, the RAISE Act, by overwhelming margins, and Governor Kathy Hochul signed it in December 2025.[65] Like California's law, the RAISE Act requires developers to annually publish a safety and security protocol describing their steps to prevent critical harms, unauthorized access, and misuse.[66] AI developers must report safety incidents associated with their models, such as behavior unaligned with user requests, unauthorized release or access, critical failure of technical or administrative controls, or unauthorized use.[67] These reporting requirements are fairly light in terms of their regulatory effect, but they may constitute important preliminary steps toward more comprehensive AI safety regulation.[68]

---

[61] *Autonomous Vehicles | Self-Driving Vehicles Enacted Legislation*, NAT'L CONF. OF STATE LEGISLATURES (Feb. 18, 2020), https://www.ncsl.org/transportation/autonomous-vehicles.

[62] Garrison Lovely, *Inside Tech's Risky Gamble to Kill State AI Regulations for a Decade*, OBSOLETE (June 29, 2025), https://www.obsolete.pub/p/inside-techs-risky-gamble-to-kill.

[63] Justine Gluck, *The Raise Act vs. SB 53: A Tale of Two Frontier AI Laws*, FUTURE OF PRIV. F. (Jan. 8, 2026), https://fpf.org/blog/the-raise-act-vs-sb-53-a-tale-of-two-frontier-ai-laws/.

[64] *Id.*

[65] Jennifer Johnson, et al., *New York Legislature Passes Sweeping AI Safety Legislation*, GLOBAL POLICY WATCH (June 24, 2025), https://www.globalpolicywatch.com/2025/06/new-york-legislature-passes-sweeping-ai-safety-legislation.

[66] *Id.*

[67] *Id.* Prior to deploying frontier models, developers are also required to publish a transparency report describing their models and summarizing the developers' assessments of catastrophic risks associated with their models. Gluck, *supra* note 63.

[68] Johnson, et al., *supra* note 65. Developers would be prohibited from deploying frontier models that create an unreasonable risk of critical harm. *Id.*

C. The Drivers of Deregulation

This section explores the reasons behind the deregulatory turn in AI policy and the difficulty of enacting meaningful regulation.[69] First, it details widespread concerns about regulation hindering the United States in a race against China to lead the world in AI technology. Second, it examines the political influence of the AI industry and the tech industry in general.

1. China and the AI Race

An important source of U.S. policymakers' reluctance to regulate AI involves concerns about losing the "AI race" to China, whose AI industry is currently second only to America's.[70] The general idea behind the AI race is that nations at the forefront of AI development are likely to realize more of the potentially large economic benefits of worldwide AI adoption.[71] AI is also expected to revolutionize military technology and strategy, and the country with the most advanced AI-based weapons is, the argument goes, likely to obtain a significant military advantage over nations without such weapons.[72] Articulating a strong version of the AI race concept, Senate Minority Leader Chuck Schumer has warned, "if America falls behind China on AI, we will fall behind everywhere: economically, militarily, scientifically, educationally."[73] The Trump Administration has adopted the AI race as an organizing principle of American AI policy. The Administration's AI Action Plan setting out its vision for AI Policy is titled "Winning the Race."[74] It sets "dominance" in the "race to achieve global dominance in [AI]" as its primary goal.[75]

---

[69] *Poll: Californians Support Strong Version of SB1047, Disagree With Anthropic's Proposed Changes*, AIPI (last visited___), https://theaipi.org/april-voters-prefer-ai-regulation-over-self-regulation-2-2/.

[70] *See, e.g.*, INTERESTING TIMES, *supra* note 34.

[71] Sherman, *supra* note 34.

[72] *E.g.*, *Remarks by Secretary Esper at National Security Commission on Artificial Intelligence Public Conference*, U.S. DEP'T OF WAR (Nov. 5, 2019) [hereinafter Esper Remarks], https://www.war.gov/News/Transcripts/Transcript/Article/2011960/remarks-by-secretary-esper-at-national-security-commission-on-artificial-intell.

[73] *With China's DeepSeek, US tech fears red threat*, THE CHRONICLE (Jan. 30, 2025), https://www.thechronicle.com.au/news/breaking-news/with-chinas-deepseek-us-tech-fears-red-threat/news-story/3c0bffe6ee0f6f54ea85c52bafbd6523.

[74] WINNING THE RACE, *supra* note 3.

[75] Id.

The concept of a great AI race against China is a powerful disincentive for domestic AI regulation. If AI regulation might restrict the private sector's progress in AI development, it could give China an advantage in the race to dominate AI.[76] Accordingly, the concept of a race against China is frequently invoked as an argument against regulation of America's AI industry. Announcing a Senate hearing on AI in May 2025, Senator Ted Cruz emphasized that "the way to beat China in the AI race is to outrace them in innovation, not saddle AI developers with European-style regulations."[77] J.D. Vance has raised concerns that, if America pauses AI development in order to focus on safety issues, we could "find ourselves all enslaved to P.R.C.-mediated A.I."[78] The CEO of a tech industry trade organization argued that the proposed 10-year moratorium on state regulation of AI was "important in the race against China."[79] A tech lobbyist made similar arguments, asserting "[w]e're in a race with China to lead the future of artificial intelligence. A fractured, state-by-state system will only slow us down."[80] Ads in favor of the moratorium argued that "we can't let China get the upper hand."[81] Concerns about ceding America's lead in AI technology to China are a major impediment to domestic AI regulation.[82]

2. The Tech Industry and the Tech Lobby

The tech industry is a major component of the American economy, and the flourishing of U.S. tech companies has contributed to the remarkable growth of America's GDP relative to Europe over the last

---

[76] *E.g.*, Edelman et al., *supra* note 34.

[77] *Winning the AI Race*, *supra* note 34.

[78] INTERESTING TIMES, *supra* note 34.

[79] William Gallagher, *USA: Big Tech allegedly pushes for 10-year ban on state AI regulation*, BUS. AND HUM. RTS. CTR. (June 18, 2025), https://www.business-humanrights.org/en/latest-news/usa-big-tech-allegedly-pushes-for-10-year-ban-on-state-ai-regulation.

[80] Owen Dahlkamp, *Ted Cruz wants to stop states from regulating AI. Some of his Republican colleagues aren't so sure*, TEX. TRIB. (June 27, 2025), https://www.texastribune.org/2025/06/27/ted-cruz-ai-moratorium-reconciliation-republicans-congress-trump.

[81] Lovely, *supra* note 62.

[82] It bears noting that the Trump administration sometimes appears to place other considerations above U.S.-China competition in policy areas affecting AI development. *See, e.g.,* Tripp Mickle, *Nvidia Says U.S. Has Lifted Restrictions on A.I. Chip Sales to China*, N.Y. TIMES (July 14, 2025), https://www.nytimes.com/2025/07/14/technology/nvidia-ai-chip-sales-china.html.

thirty years.[83] AI companies in particular have undergone remarkable growth and received billions of investment dollars in recent years, and Nvidia, the primary maker of the computer chips that enable frontier AI models, is now the largest U.S. company by market capitalization.[84] Many arguments against substantial regulations of AI companies are centered around their potential for generating economic prosperity.[85] The EU's AI Act is often invoked as evidence of how aggressive regulation can stifle the AI industry in a region.[86] Along these lines, policymakers have expressed concern for protecting the growth of AI companies, promoting AI development, and avoiding overregulation of a relatively new industry.[87]

AI companies' lobbying efforts have also grown dramatically in recent years, and the tech lobby is largely opposed to meaningful AI regulation.[88] AI companies and venture capital firms have attempted to block or water down numerous proposed laws addressing AI, with considerable success.[89] Their lobbyists are "very visible [and] present on the hill," as they nurture relationships with senators and representatives in Congress.[90] "The overarching ask is for no regulation or for light-touch regulation, and so far, they've gotten that."[91]

---

[83] *See, e.g.*, Paul Krugman, *Why Does U.S. Technology Rule?*, SUBSTACK (Dec. 12, 2024), https://paulkrugman.substack.com/p/why-does-us-technology-rule; FREDERIK ERIXON, OSCAR GUINEA & OSCAR DU ROY, IF THE EU WAS A STATE IN THE UNITED STATES: COMPARING ECONOMIC GROWTH BETWEEN EU AND US STATES (2023), https://ecipe.org/wp-content/uploads/2023/06/ECI_23_PolicyBrief_07-2023_LY02.pdf.

[84] *E.g.*, Hayden Field & MacKenizie Sigalos, *AI craze is distorting VC market, as tech giants like Microsoft and Amazon pour in billions of dollars*, CNBC (Sep. 6, 2024), https://www.cnbc.com/2024/09/06/ai-craze-getting-funded-by-tech-giants-distorting-traditional-vcs.html.

[85] *See, e.g.*, Oliver Roberts, *EU AI Act's Burdensome Regulations Could Impair AI Innovation*, BLOOMBERG LAW (Feb. 21, 2025), https://news.bloomberglaw.com/us-law-week/eu-ai-acts-burdensome-regulations-could-impair-ai-innovation.

[86] *See, e.g.*, *id.*

[87] *See, e.g.*, Thomas Adamson & Aamer Madhani, *Vance rails against 'excessive' regulation at Paris AI summit*, PBS NEWS (Feb. 11, 2025), https://www.pbs.org/newshour/politics/vance-rails-against-excessive-regulation-at-paris-ai-summit; Letter from Governor Gavin Newsom Accompanying Veto of S.B. 1047 (Sep. 29, 2024), https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf;

[88] Mohar Chatterjee, *The AI lobby plants its flag in Washington*, POLITICO (June 6, 2025), https://www.politico.com/news/2025/06/06/the-ai-lobby-plants-its-flag-in-washington-00389549.

[89] *See, e.g.*, Lovely, *supra* note 62.

[90] Chatterjee, *supra* note 88 (quoting representative Don Beyer (D-Va)).

[91] *Id.* (quoting Doug Calidas, senior vice president of government affairs for the AI policy nonprofit Americans for Responsible Innovation).

In the summer of 2025, executives from OpenAI and pro-AI venture capital firm Andreesen Horowitz set up a super-PAC with more than $100 million in funding to advocate against AI regulations.[92] Around the same time, Meta launched two super-PACs geared toward fighting AI regulation at the state level.[93] These AI super-PACs are likely to be among the largest spenders in the 2026 midterm elections.[94] At all levels of government, policymakers have been reluctant to regulate AI companies because those firms do not want to be regulated, and because they wield substantial economic and lobbying clout.

II. THE EXISTENTIAL RISK DEBATE

The recognition of AI existential risk has migrated from science-fiction conventions to the world's major newspapers and New York Times Best Seller Lists.[95] Nobel laureates, leading engineers, and the heads of major AI laboratories now openly discuss the possibility that AI could pose catastrophic risks to humanity.[96] Indeed, these risks have become so widely acknowledged by industry leaders that skeptics have begun to wonder whether they might be exaggerating the dangers of their technology in some complex attempt to foster regulation that would benefit incumbents—though at every juncture, these same firms and their lobbying arms oppose actual meaningful regulation.[97]

On the merits of this debate, a threshold question looms large: who must prove what, and at what level of proof? Skeptics treat extinction claims as extraordinary and therefore demand extraordinary

---

[92] Amrith Ramkumar & Brian Schwartz, *Silicon Valley Launches Pro-AI PACs to Defend Industry in Midterm Elections*, WALL ST. J. (Aug. 25, 2025), https://www.wsj.com/politics/silicon-valley-launches-pro-ai-pacs-to-defend-industry-in-midterm-elections-287905b3?st=beGTwg&reflink=desktopwebshare_permalink.

[93] Eli Tan & Theodore Schleifer, *Meta Ramps Up Spending on A.I. Politics With New Super PAC*, N.Y. TIMES (Sep. 23, 2025), https://www.nytimes.com/2025/09/23/technology/meta-super-pac-ai.html.

[94] *See id.*

[95] *See supra* note 17; *Best Seller, Combined Print & E-Book Nonfiction*, N.Y. TIMES (Oct. 5, 2025), https://www.nytimes.com/books/best-sellers/2025/10/05/combined-print-and-e-book-nonfiction (listing ELIEZER YUDKOWSKY & NATE SOARES, IF ANYONE BUILDS IT, EVERYONE DIES (2025)).

[96] *See, e.g.*, *supra* note 15; Matteo Wong, *Anthropic Is at War with Itself*, ATLANTIC (Jan. 28, 2026, 4:44 pm ET), https://www.theatlantic.com/technology/2026/01/anthropic-is-at-war-with-itself/684892.

[97] *See supra* Part I.C.2.

evidence.[98] Until bodies pile up—actual mass harm, not an Uber fatality or a flash crash[99]—they say the claim remains speculative. Proponents reply that "safe until proven dangerous" is itself an extraordinary default for technologies that act autonomously, can solve problems with superhuman speed, and offer no ex-post *undo* button.[100] Regulatory history, they argue, is written in blood: cyber-hacks, reactor meltdowns, environmental disasters, and market seizures were all speculative the day before they happened.[101] This is not to say that potentially dangerous modern technologies should not have been developed, but rather that safe-by-default is the wrong frame for thinking of transformative technologies.

Into this stalemate concerning AI's future harms, we bring a lawyer's toolkit: assessing evidentiary sufficiency, allocating burdens of proof, and deriving optimal regulatory policies under uncertainty. Our goal is to assess whether current evidence warrants regulation, and if so, when, at what intensity, and through what mechanisms.

Before proceeding, one premise requires emphasis. The public image of AI, a chatbot answering questions in a browser window, easily closed with a click, is increasingly incongruent with reality. Frontier AI systems are becoming *agentic*: given an objective, they decompose it into sub-tasks, select tools, execute multi-step plans, and adapt when obstacles arise.[102] They can already complete tasks reliably that would occupy a skilled human for hours, and that capability horizon is doubling roughly every seven months.[103] This agentification transforms the regulatory question, as policymakers are increasingly

---

[98] *See, e.g.*, Dylan Matthews, *Why can't anyone agree on how dangerous AI will be?*, Vox (Mar. 13, 2024), https://www.vox.com/future-perfect/2024/3/13/24093357/ai-risk-extinction-forecasters-tetlock.

[99] A flash crash is a "very rapid, deep, and volatile fall in security prices occurring within an extremely short time period," typically followed by a quick recovery. *Flash Crash*, INVESTOPEDIA, https://www.investopedia.com/terms/f/flash-crash.asp (last visited Feb. 11, 2026).

[100] *See generally* Dan Hendrycks et al., *An Overview of Catastrophic AI Risks* (Oct. 9, 2023), https://arxiv.org/pdf/2306.12001.pdf.

[101] *See, e.g.*, Brian Judge et al., *When Code Isn't Law: Rethinking Regulation for Artificial Intelligence*, 44 POL'Y AND SOC'Y 85, 88 (2025); Laura Gil, *Learning from Fukushima Daiichi: Factors Leading to the Accident*, 62 IAEA Bull. (Mar. 2021), https://www.iaea.org/bulletin/learning-from-fukushima-daiichi-factors-leading-to-the-accident; Jill Treanor, *The 2010 'Flash Crash': How it Unfolded*, GUARDIAN (Apr. 22, 2015), https://www.theguardian.com/business/2015/apr/22/2010-flash-crash-new-york-stock-exchange-unfolded; Brenda Eskenazi, et al., *The Pine River Statement: Human Health Consequences of DDT Use*, 117 ENVIRON. HEALTH PERSPECT. 1359 (2009).

[102] Miller, *supra* note 10.

[103] Kwa et al., *supra* note 11.

tasked with governing systems that act in real-world contexts, from vehicles to supply chains.[104]

We structure our analysis below around three intersecting risk pathways: human-directed harms (where AI amplifies malicious intent), accident risks (where autonomous systems fail catastrophically), and loss-of-control scenarios (where advanced AI pursues objectives misaligned with human values).

A. Human Directed Risks

Perhaps the most immediate and well-documented category of AI existential risk stems from a familiar source: human action. Here the risk profile does not depend on unexpected system failure or loss of control, but rather, with systems that perform *exactly as told*. If human operators direct these systems to accomplish bad social ends, for reasons of greed, malice, jingoism, racism, or any other human motivation, then the scale of risk directly scales with the power of these systems. As AI becomes stronger, so do its destructive capabilities.

The core danger of agentic AI in malicious hands lies in how it inverts traditional security assumptions. Conventional defense presumes that complex attacks require proportionally complex planning, coordination, and expertise from attackers.[105] For example, a biological weapons program requires trained scientists, a long supply chain, and generous funding.[106] Likewise, a sophisticated cyber intrusion demands skilled hackers with physical access to infrastructure or the ability to social engineer.[107] Agentic AI flattens these hurdles, enabling attackers with minimal technical sophistication to command operations that would previously require teams of specialists. The technology offers three critical dimensions that can amplify destructive capabilities: reducing the expertise

---

[104] *See Untapped Edge*, DELOITTE (Jan. 21, 2026) https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html.

[105] *See* Brandon del Pozo, *Complexity, Lethality, and the Perverse Imagination: Modelling Nonstate Actors' Means of Attack*, 46 STUD. CONFL. & TERRORISM 13 (2023).

[106] Stacey L. Knobler et al., Biological Threats and Terrorism: Assessing the Science and Response Capabilities: Workshop Summary 211 (2002), https://www.nationalacademies.org/read/10290/chapter/1.

[107] *See, e.g.*, VERIZON, 2023 DATA BREACH INVESTIGATIONS REPORT 46 (2023), https://www.verizon.com/business/resources/reports/dbir.

required for complex attacks, enabling unprecedented scale of operations, and ensuring persistence that outlasts human actors.

1. Expertise

Where a traditional hacker might spend months planning a cyberattack, an agentic AI system can receive high-level commands like "compromise this 'organization's financial systems" or "coordinate denial-of-service attacks on these infrastructure targets" and then formulate and execute the necessary subgoals autonomously. Indeed, a recent report finds evidence of such agentic cyberattacks on American companies, likely orchestrated by a foreign state actor.[108] This represents more than efficient information retrieval; it is a qualitative "uplift" in capability that security experts are only beginning to grapple with.[109]

While cyberattacks may never threaten truly irreversible harm, expertise can facilitate other, potentially more existential attacks. Facilitating the creation of bioweapons by bridging gaps in expertise may lead to irreversible harms.[110] However, skeptics contend that the main bottlenecks are not really ones of knowledge.[111] There are plenty of books on pathogens and biology, and Google can locate much of this material.[112] The real bottlenecks are much more physical, consisting of access to samples, reagents, and wetware, alongside significant "soft skills" that AI lacks.[113] Risk proponents reply that agentic AI could soon provide significant uplift even in material sourcing and handling.[114] While the degree of risk is subject to debate, increasing AI expertise across a variety of fields lowers costs for attackers and increases the capabilities of state and non-state actors to cause harm.

---

[108] MICROSOFT, DIGITAL DEFENSE REPORT 2025 49 (2025), https://www.microsoft.com/en-us/security/business/digital-defense-report.

[109] *See GPT-4 System Card*, OPENAI 52 (2023), https://cdn.openai.com/papers/gpt-4-system-card.pdf.

[110] *E.g.*, NAT'L ACADS. OF SCIS., ENG'G, & MED., BIODEFENSE IN THE AGE OF SYNTHETIC BIOLOGY 6 (2018), https://doi.org/10.17226/24890.

[111] *See id.*

[112] *See id.* at 96.

[113] *See id.* at 6.

[114] *See* David Luckey et al., *Mitigating Risks at the Intersection of Artificial Intelligence and Chemical and Biological Weapons*, HOMELAND SECURITY OPERATIONAL ANALYSIS CTR. (2025), available at https://www.rand.org/pubs/research_reports/RRA2990-1.html.

2. Scale

The second force multiplier, scale, fundamentally alters threat dynamics. A single individual can direct thousands of autonomous agents simultaneously, each pursuing complex objectives that previously required dedicated human operators. This capability is already visible in AI-powered bot networks that attempt to manipulate public opinion and financial markets through coordinated, high-volume activity.[115] Where conventional information warfare requires teams of operators to maintain long-running campaigns across multiple platforms, a single actor with access to agentic AI could orchestrate bot armies that degrade the shared information environment that democratic deliberation requires.[116]

Scale advantages defenders as well. A single defender can deploy swarms of monitoring systems that identify distributed attack patterns invisible to human analysts.[117] AI-powered cybersecurity tools, automated threat detection systems, and predictive defense algorithms are already being developed and deployed.[118] If AI can help attackers find vulnerabilities, the argument goes, it can equally help defenders identify and patch them.[119]

Yet several asymmetries favor attackers even as both sides gain capabilities. First is the fundamental rule of security imbalance: defenders must secure all potential vulnerabilities while attackers need exploit only one.[120] This asymmetry compounds as system complexity increases.[121] Second, identifying suspicious activity becomes a needle-in-haystack problem as AI agents proliferate; most agentic behavior will be entirely innocent, allowing attackers to hide hostile operations within the noise of ordinary automated transactions. Third, economics often favor offense: crafting an effective attack can be done by a small team with appropriate tools,

---

[115] Tom C. W. Lin, *AI and the Future of Market Manipulation*, 85 OHIO ST. L.J. , 685, 697–700 (2024).

[116] Daniel Thilo Schroeder et al., *How Malicious AI Swarms Can Threaten Democracy* (Jan. 22, 2026), https://arxiv.org/abs/2506.06299.

[117] Samaresh Kumar Singh & Joyjit Roy, *Decentralized Multi-Agent Swarms for Autonomous Grid Security in Industrial IoT: A Consensus-based Approach*, (Jan. 24, 2026), https://www.arxiv.org/pdf/2601.17303.

[118] *Id*. at 33.

[119] *See* Andrew J. Lohn, *The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict*, at 63 (Apr. 17, 2025), https://arxiv.org/abs/2504.13371.

[120] *Id*. at 27.

[121] *Id*. at 34.

while sophisticated defense generally requires substantial resources and coordination across multiple organizations, even in the era of defensive AI agents.[122] Fourth, the temporal dimension favors attackers, who can choose the timing of their operations and probe defenses until they identify a moment of weakness.[123]

We cannot yet know whether AI will ultimately favor attack or defense in the long run. But the catastrophic potential of a single successful, AI-augmented attack counsels against banking on a purely optimistic outcome. Under deep uncertainty about whether defensive applications can keep pace with offensive capabilities, hoping that defense will naturally prevail represents wishful thinking rather than serious risk management.

3. Persistence

The third force multiplier, persistence, may prove most consequential because it breaks a fundamental assumption in security thinking. Traditional threats are inherently time limited. Attackers get arrested, distracted, or simply lose interest. Criminal enterprises require ongoing coordination to maintain operations.[124] Terrorist cells fracture when key members are eliminated.[125] The assumption, deeply embedded in security doctrine, is that stopping the humans stops the threat.[126]

Agentic AI breaks this assumption. An attacker who deploys autonomous agents with conditional logic can create operations that continue indefinitely without further human involvement. The agents require no salary, feel no fear, and never lose interest. They can lie dormant for months or years, waiting for specified conditions ("hooks")

---

[122] William A. Owens, Kenneth W. Dam & Herbert S. Lin, *Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities*, NAT'L ACAD. OF SCIS. 5 (2009). https://sites.nationalacademies.org/cs/groups/cstbsite/documents/webpage/cstb_050541.pdf.

[123] Lohn, *supra* note 119, at 63.

[124] *Id.* at 24.

[125] Bryan C. Price, *Leadership Decapitation and the End of Terrorist Groups*, BELFER CTR. FOR SCI. & INT'L AFFS., HARV. KENNEDY SCH. (May 2012), https://www.belfercenter.org/publication/leadership-decapitation-and-end-terrorist-groups.

[126] *See id.*

before activating.[127] They can adapt to countermeasures and continue pursuing objectives long after their creator is dead or in prison.

This persistence is especially problematic for legal enforcement. Traditional interdiction focuses on stopping or detaining human actors. These approaches presume ongoing human involvement in executing attacks, a presumption that persistent AI agents may wholly or partially undermine. Attribution and intent may also be more difficult to establish in contexts where AI agents replace human co-conspirators. Determining who designed an AI agent, or what they directed the agent to do as opposed to what the agent did on its own, could be virtually impossible, complicating attribution and mens rea determinations.

4. Conflict

Military applications of agentic AI are perhaps the clearest and most concerning example of human-directed AI risks. The integration of AI into weapons systems is no longer experimental. It represents operational doctrine, driven by perceived strategic necessity and substantial defense investment.[128] The trajectory of military practice underscores how rapidly this landscape is shifting. Around the world, armed forces are embedding AI throughout their operations, including live-fire use in ongoing conflicts.[129] In late 2025 alone, the U.S. Department of Defense launched an AI platform for roughly three million personnel; Germany approved six hundred million euros in AI-enabled loitering munitions; the United Kingdom signed a £1.5 billion contract for AI-driven targeting systems tested in Ukraine; Israel reorganized its command structure around a dedicated AI division; and China publicly showcased autonomous combat systems at a national military parade.[130] AI-enabled weapons capable of

---

[127] *See* Georgia Collins, *AI Agents Drive First Large-Scale Autonomous Cyberattack*, CYBERMAGAZINE (Oct. 16, 2024), https://cybermagazine.com/news/ai-agents-drive-first-large-scale-autonomous-cyberattack; MJ Kubba, *Automate your development workflow with Kiro's AI agent hooks*, KIRO (July 16, 2025), https://kiro.dev/blog/automate-your-development-workflow-with-agent-hooks.

[128] *E.g.*, Hegseth Memorandum, *supra* note 8, at 1.

[129] *See infra* notes 156–158.

[130] Press Release, U.S. Dep't of War, The War Department Unleashes AI on New GenAI.mil Platform (Dec. 9, 2025), https://www.war.gov/News/Releases/Release/Article/4354916/the-war-department-unleashes-ai-on-new-genaimil-platform; Haye Kesteloo, *Helsing And Stark Beat*

autonomously identifying and engaging targets have already been field-tested in Ukraine, Russia, the Middle East, and other theaters.[131] According to the UN, Turkish-made Kargu-2 loitering munitions deployed in Libya can "hunt down and remotely engage" retreating forces "without requiring data connectivity between the operator and the munition," suggesting that autonomous killing machines are already in use.[132] The United States has taken steps to partially integrate AI into its nuclear defense systems,[133] while the U.N. Secretary-General warns that military AI "exacerbates the existential risks posed by nuclear weapons" and heightens "the risks of inadvertent escalation and miscalculation."[134]

What equilibrium emerges from these capabilities is contested. Skeptics emphasize that the same technologies that empower attackers also strengthen defenders.[135] Risk proponents stress a different dynamic: proliferation.[136] AI capabilities are intrinsically dual-use, the methods for training and fine-tuning frontier models are documented in the open literature, and the underlying hardware

---

Rheinmetall For €600 Million German Loitering Munition Contracts, DRONEXL (Jan. 27, 2026), https://dronexl.co/2026/01/27/helsing-stark-rheinmetall-loitering-munition; Press Release, U.K. Ministry of Defense, *New strategic partnership to unlock billions and boost military AI and innovation* (Sept. 18, 2025), https://www.gov.uk/government/news/new-strategic-partnership-to-unlock-billions-and-boost-military-ai-and-innovation; David Michael Swindle, *The Future of War: Israel Takes Global Lead on Military Innovation With New AI Division, Iron Beam Laser System*, ALGEMEINER (Dec. 4, 2025), https://www.algemeiner.com/2025/12/04/future-war-israel-takes-global-lead-military-innovation-new-ai-division-iron-beam-laser-system; Brandi Vincent, *China's military technology parade underscores need for more U.S. deterrents, experts say*, DEFENSESCOOP (Sept. 5, 2025), https://defensescoop.com/2025/09/05/china-military-technology-victory-day-parade-u-s-analysts.

[131] *See, e.g.*, Kateryna Stepanenko, *The Battlefield AI Revolution Is Not Here Yet: The Status of Russian and Ukrainian AI Drone Efforts*, INST. FOR THE STUDY OF WAR (June 2, 2025), https://understandingwar.org/research/russia-ukraine/the-battlefield-ai-revolution-is-not-here-yet-the-status-of-current-russian-and-ukrainian-ai-drone-efforts; Daniel Estrin, *After two years of war, Israeli weapons makers showcase their new tech*, NPR (Dec. 3, 2025), https://www.npr.org/2025/12/03/nx-s1-5628404/after-two-years-of-war-israeli-weapons-makers-showcase-their-new-tech.

[132] MAJUMDAR ROY CHOUDHURY, ET AL., FINAL REPORT OF THE PANEL OF EXPERTS ON LIBYA, U.N. Doc. S/2021/229 (Mar. 8, 2021), https://www.securitycouncilreport.org/un-documents/document/s-2021-229.php.

[133] Greg Hadley, *AI 'Will Enhance' Nuclear Command and Control, Says STRATCOM Boss*, AIR & SPACE FORCES MAG. (Oct. 28, 2024), https://www.airandspaceforces.com/stratcom-boss-ai-nuclear-command-control.

[134] U.N. SECRETARY-GENERAL, ARTIFICIAL INTELLIGENCE IN THE MILITARY DOMAIN, U.N. Doc. A/80/78 (June 5, 2025).

[135] *See* Lohn, *supra* note 119, at 71.

[136] *See supra* note 100.

(drones, for example) is often widely available.[137] Those features can democratize access to extremely dangerous weapons. Deterrence becomes fragile as the number of actors grows.

The development of AI weapons also creates pathways to strategic instability that extend beyond proliferation concerns. When military action can be conducted through autonomous systems rather than human soldiers, the political costs of military operations decline dramatically.[138] While these wars might reduce direct human casualties, the ensuing environmental harm and political repercussions could be severe.[139] In addition, the speed of AI-enabled military operations may exceed human decision-making capabilities, creating what strategists call "time-domain compression."[140] When autonomous systems can observe, orient, decide, and act at machine speed, the window for human deliberation shrinks from hours or days to minutes or seconds. This temporal compression interacts dangerously with both conventional military deployment and nuclear command systems.

B. Accidental and Systemic Risk

The previous section examined risks from deliberate weaponization. Yet catastrophic outcomes need not involve malicious intent. Well-meaning actors deploying AI systems for beneficial purposes face serious risks from accidents, misalignment, and cascading failures. These risks differ fundamentally from human-directed threats because they emerge not from hostile design but from the interaction between autonomous decision-making and environmental complexity. What makes accidental risks particularly concerning is their potential to become systemic. When AI systems integrate deeply into interconnected infrastructure, individual

---

[137] *See* Daniel Kang et al., *Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks*, (Feb. 11, 2023), https://arxiv.org/abs/2302.05733.

[138] Paul J. Springer, *Military Robotics Might Enable Conflict While Reducing Costs*, FOREIGN POLICY RESEARCH INSTITUTE (Apr. 23, 2019), https://www.fpri.org/article/2019/04/military-robotics-might-enable-conflict-while-reducing-costs.

[139] *Id.*

[140] Deb Henley, *Human-Machine Teaming in Battle Management: A Collaborative Effort Across Borders*, U.S. AIR FORCE, NELLIS AIR FORCE BASE (Jan. 5, 2026), https://www.nellis.af.mil/News/Article/2003850932/human-machine-teaming-boosts-battle-management-speed-and-accuracy.

accidents can trigger cascades that propagate across multiple domains faster than human operators can recognize or respond.

1. Infrastructure

Modern infrastructure networks exhibit complex interdependencies that amplify failure propagation. AI integration into these networks increases the risk of potentially catastrophic failures. The scope of AI integration into America's power grid, to take one example, is accelerating rapidly. Research by the consultancy Gartner predicts AI adoption in forty percent of power and utilities control rooms by 2027, and ninety-four percent of power and utility CIOs planned to increase their AI investments in 2025 by an average of 38.3%.[141] Ten to fifteen percent of water utilities deploy AI for treatment optimization, and that proportion is expected to rapidly increase.[142] Financial institutions manage trillions of dollars in assets through AI systems.[143] As AI becomes ubiquitous across critical systems, the line between isolated accidents and systemic failures begins to blur.

AI systems can fail catastrophically when encountering conditions outside their training data, what machine learning researchers term "out-of-distribution" environments.[144] Risk compounds when multiple AI systems share similar architectures, training data, or operating assumptions. Market concentration amplifies this concern. Varied AI applications are likely to use the same leading base model architectures, meaning a fundamental flaw or adversarial attack could trigger failures across systems that appear independent.

This architectural homogeneity creates correlated failure modes that defeat the independence assumptions underlying traditional risk management. When diverse systems fail independently, redundancy and backup protocols can provide effective safeguards. When systems fail simultaneously due to shared vulnerabilities, those safeguards can collapse.

---

[141] *Gartner Predicts*, *supra* note 12.
[142] SMART WATER, *supra* note 12.
[143] *See* O'Connell, *supra* note 12.
[144] See Andrew J. Lohn, *Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-of-Distribution Performance*, at 5-6 (Sep. 2, 2020), https://arxiv.org/pdf/2009.00802.

2. Goal specification and Goodhart's Law

Even when AI systems operate within their training distribution, they may pursue assigned objectives through unanticipated means that cause significant harm. This phenomenon, often called "specification gaming," reflects Goodhart's Law: when a specified goal becomes the target of optimization, the system may find ways to achieve it that undermine the reason we cared about the goal in the first place.[145] Put more concretely, Goodhart problems arise when AI systems pursue specified goals in ways their developers did not intend and that thwart or harm their developers' interests.

The boat racing game provides an early example. OpenAI researchers trained an AI agent to achieve a high score in a racing video game, only to discover that it learned to spin in place endlessly, accumulating bonus points rather than finishing the race.[146] The system maximized its reward function while subverting the intended behavior.[147] Similar patterns appear across decades of AI research, from a tic-tac-toe AI that won by causing opponents' programs to crash to the GenProg system that "fixed" bugs by deleting the test files that detected them.[148]

The advent of LLMs has made skeptics doubt the practical significance of such problems. LLMs demonstrate ability to respond not just to users' literal requests but also to their perceived intentions. Because frontier models are trained through extensive reinforcement learning from human feedback (RLHF), they are demonstrably capable of learning broader human goals.[149] And anyone who has used a modern chatbot can attest that it usually does what is meant, not merely what is said.

These improvements are real and should not be dismissed. But safety researchers offer several grounds for doubting whether this apparent progress translates into the kind of reliability that deployment in critical systems would require. The core issue is one of

---

[145] Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, DEEPMIND SAFETY RESEARCH (Apr. 21, 2020), https://deepmindsafetyresearch.medium.com/specification-gaming-the-flip-side-of-ai-ingenuity-c85bdb0deeb4.

[146] *Faulty Reward Functions in the Wild*, OPENAI (Dec. 21, 2016), https://openai.com/blog/faulty-reward-functions.

[147] *Id.*

[148] Krakovna et al., *supra* note 145.

[149] Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, 35 ADVANCES IN NEURAL INFO. PROCESSING SYS. 27730 (2022).

generalization. During training, LLMs are shown many examples of what humans prefer and dislike, and they learn to generalize from these examples to novel cases.[150] Because training corpora are vast and diverse, models appear to act as expected in familiar situations.[151] But it is far from clear that in genuinely novel situations, precisely where specification failures matter most, the model will have learned the correct lessons from its RLHF interactions.

One vivid demonstration is the problem of sycophancy. Models systematically tell users what they want to hear because, in training data, agreement correlates with higher ratings.[152] Empirically, models exposed to user opinions agreed with demonstrably incorrect beliefs in a majority of cases, with susceptibility *increasing* with model scale.[153] This is Goodharting at a higher level of abstraction: the model optimizes for the appearance of helpfulness rather than helpfulness itself. More troubling, Anthropic's "Sycophancy to Subterfuge" research demonstrates that AI models can start with simple sycophancy and then generalize to tampering with their own evaluation functions and, in some cases, covering their tracks.[154]

A second concern is that even within their current training distributions, LLMs exhibit specification gaming at non-trivial rates. Anthropic's 2025 research on reward hacking in realistic coding environments found that RLHF-trained models learned to pass tests by calling `sys.exit(0)` to fake success, a trick analogous to a student claiming to pass an exam by walking out before the proctor could mark any answers wrong.[155] Production models exhibited such reward-hacking behaviors in 12% to 18% of relevant cases.[156] More vividly, Replit's coding agent, in late 2025, deleted an entire production database containing over 1,200 records against explicit commands,

---

[150] Steven Thomas, *Do Large Language Models Really Generalize?*, MEDIUM (June 5, 202.4), https://nyudatascience.medium.com/do-large-language-models-really-generalize-this-paper-says-yes-ca99fbb00a44

[151] *See supra* note 149.

[152] Mrinank Sharma et al., *Towards Understanding Sycophancy in Language Models*, ANTHROPIC (Oct. 23, 2023), https://www.anthropic.com/news/towards-understanding-sycophancy-in-language-models.

[153] *Id.*

[154] Carson Denison et al., *Sycophancy to Subterfuge: Investigating Reward Tampering in Language Models*, ANTHROPIC (2024), https://www.anthropic.com/research/reward-tampering.

[155] Monte MacDiarmid et al., *From Shortcuts to Sabotage: Natural Emergent Misalignment from Reward Hacking*, ANTHROPIC (Nov. 21, 2025), https://www.anthropic.com/research/emergent-misalignment-reward-hacking.

[156] *Id.*

later describing its actions as "a catastrophic error in judgment."[157] Similar incidents involving coding agents deleting entire codebases have been documented across multiple platforms.[158]

These concerns should be understood in regulatory rather than absolute terms. The question is not whether LLMs are useful, because they plainly are, but whether they meet the reliability standards appropriate for systems increasingly embedded in critical infrastructure. Engineers speak of the "march of nines": the difference between 99% reliability and 99.99% reliability, and the even higher level expected of systems where failures are catastrophic. Current LLMs offer no such assurances. Indeed, we lack validated methods for measuring their reliability at the tails of their behavioral distributions.[159] For systems that will increasingly control supply chains and manage power grids, this uncertainty is itself a regulatory problem.

3. Speed, scale, and systemic failure

One might expect that deployers of AI in critical infrastructure would implement safeguards before deployment, and that accidents, even serious ones, would serve as canaries in the coal mine, allowing operators to shut down failing systems before harm propagates. As with nuclear meltdowns like the Three Mile Island incident, human operators could respond to accidents and override critical systems to avert further catastrophe.[160]

The concern with AI is that the combination of speed and scale fundamentally alters risk profiles in ways that traditional safety mechanisms cannot address. Trading algorithms execute thousands of transactions in microseconds.[161] Fleets of autonomous vehicles make

---

[157] Bryan Reynolds, *The Replit AI Disaster: A Wake-Up Call for Every Executive on AI in Production*, BAYTECH (July 23, 2025), https://www.baytechconsulting.com/blog/the-replit-ai-disaster-a-wake-up-call-for-every-executive-on-ai-in-production.

[158] *See, e.g.*, Vivek Athreya, *How an AI Agent Deleted My Entire Codebase (And What I Learned)*, MEDIUM (Sep. 7, 2025), https://medium.com/@athreya.vivek/how-an-ai-agent-deleted-my-entire-codebase-and-what-i-learned-ab23a2e7f22f; Frank Landymore, *Google's AI Deletes User's Entire Hard Drive, Issues Groveling Apology: "I Cannot Express How Sorry I Am"*, FUTURISM (Dec 6, 2025, 1:30 PM EST), https://futurism.com/artificial-intelligence/google-ai-deletes-entire-drive.

[159] *See* Sydney Von Arx, Lawrence Chan & Elizabeth Barnes, *Recent Frontier Models Are Reward Hacking*, METR (June 5, 2025), https://metr.org/blog/2025-06-05-recent-reward-hacking.

[160] *See, e.g.*, *Three Mile Island Accident*, WORLD NUCLEAR ASS'N (Oct. 11, 2022), https://world-nuclear.org/information-library/safety-and-security/safety-of-plants/three-mile-island-accident.

[161] Bill Conerly, *High Frequency Trading Explained Simply*, FORBES (Dec. 14, 2020), https://www.forbes.com/sites/billconerly/2014/04/14/high-frequency-trading-explained-simply.

steering decisions faster than human reaction time.[162] Power grid optimization systems adjust load balancing continuously across interconnected networks[163]. This speed mismatch means that by the time humans recognize a problem, AI systems may have already caused substantial harm, and the harm may have propagated to other systems before anyone understands what is happening.

Modern infrastructure is deeply interdependent in ways that may not be immediately apparent. Power grids depend on telecommunications for coordination.[164] Financial systems depend on power and telecommunications.[165] Supply chains depend on all three.[166] Water treatment depends on power, telecommunications, and supply chains.[167] Emergency services depend on the entire stack.[168] When AI systems integrate throughout this network, making real-time decisions based on conditions across multiple domains, a failure in one sector can propagate to others faster than human operators can diagnose, much less prevent.

Skeptics might respond that AI could serve as its own solution. An AI auditor could monitor critical systems in real time, identify errors, and order immediate shutdown if necessary.[169] Combined with proper

---

[162] Shengbo Wang et al., *Ultrafast Visual Perception Beyond Human Capabilities Enabled by Motion Analysis Using Synaptic Transistors*, 17 NAT. COMMUN. 1, 3 (2026).

[163] Dnyandeo Baburao Shinde, *Energy Grid Optimization- AI & Digital Technologies for Improving Efficiency*, CYIENT (Dec. 24, 2024), https://www.cyient.com/blog/energy-grid-optimization-ai-digital-technologies-for-improving-efficiency.

[164] Oliver Jung, *The Increasingly Critical Role of Communication Networks in Enhancing Power Grid Resilience Under Climate Change*, SMARTGREENS-14TH INTERNATIONAL CONFERENCE ON SMART CITIES AND GREEN ICT SYSTEMS 180, 182 (2025), https://www.scitepress.org/Papers/2025/134815/134815.pdf.

[165] Setareh Katircioglu, *The Role of Financial Systems in Energy Demand: A Comparison of Developed and Developing Countries*, 7 HELIYON e07323 (2021), https://pmc.ncbi.nlm.nih.gov/articles/PMC8225962/pdf/main.pdf.

[166] See Zlata Tabachova et al., *Estimating the Impact of Supply Chain Network Contagion on Financial Stability*, 75 J. FIN. STABILITY 1 (2024).

[167] See Jan Korzeniowski, *Complexity of Water, Wastewater Plants Creates Opportunity for Innovation and Cost Savings*, ENV'T SCI. & ENG'G MAG. (Nov. 12, 2019), https://esemag.com/water/complexity-water-wastewater-plants-creates-opportunity.

[168] *See generally* Diana Mitsova et al., *Evaluating the Impact of Infrastructure Interdependencies on the Emergency Services Sector and Critical Support Functions Using an Expert Opinion Survey*, 26 J. INFRASTRUCTURE SYS. 1 (2020).

[169] *See* Shoyab Ali, *Vision AI Based Machine Shutdowns: viAct's Role in Safety Automation*, VIACT (Oct. 31, 2025), https://www.viact.ai/post/how-viacts-vision-ai-powers-safer-machine-shutdowns?utm_source=chatgpt.com.

system isolation and redundancy, such safeguards might ensure that accidents remain contained.[170]

From a regulatory perspective, this response is insufficient for several reasons. First, the auditing AI, the system with permissions necessary to shut down critical infrastructure, will itself be susceptible to errors and adversarial manipulation. We would be solving the problem of AI reliability by introducing another AI whose reliability we must also ensure, and whose failure modes may be even harder to anticipate given its broader scope of authority. Second, it is far from clear that auditors can anticipate how shutdown of one seemingly independent system would affect other AI-driven systems. These spillover effects may be virtually impossible to anticipate, and even if we could anticipate them, it is not clear that any single system could arrest them in time. Third, and most fundamentally, engineering cannot anticipate all system faults, even when engineers have every incentive to exercise caution and abundant historical data to guide them. The 1953 North Sea flood killed 2,551 people because Dutch engineers, among the most sophisticated hydraulic engineers in the world, had designed their dams for severe-but-imaginable storms rather than 1-in-300-year events.[171] If engineers with centuries of flood data cannot reliably anticipate tail risks in a domain governed by well-understood physics, we should not expect AI system designers to anticipate all failure modes in systems whose behavior emerges from billions of parameters trained on data in ways that remain opaque even to their creators. All of these issues combine to create a risk landscape where accidental failures could trigger systemic consequences threatening the essential infrastructure on which modern civilization depends.

C. Loss of Control

The previous sections examined risks where AI systems either follow malicious human instructions or fail catastrophically while pursuing assigned objectives. A third category of risk emerges when AI systems develop and pursue goals that diverge from human intent,

---

[170] *See* Young Gyu No et al., *Monitoring Severe Accidents Using AI Techniques*, 44 NUCLEAR ENG'G & TECH. 393, 393-394 (2011) (suggesting AI could provide more capable responses to potential nuclear meltdowns than humans).

[171] PIER VELLINGA & JEROEN AERTS, NATURAL DISASTERS AND ADAPTATIONS TO CLIMATE CHANGE 136 (Sarah Boulter et al. eds., 2013).

potentially in ways that actively resist human correction. This represents a profound challenge: ensuring that increasingly autonomous systems remain aligned with human values and interests as their capabilities expand.

1. Current Challenges

Current AI systems already exhibit behaviors that reflect the challenges of aligning AI with human goals. Large language models navigate complex tensions between refusal and censorship, sometimes in ways that reveal misalignment with intended values. Grok produces non-consensual sexualized images of adults and children.[172] Frontier AI models often strategically withold or misrepresent information when doing so helps acheieve goals, like the OpenAI o3 model that reasoned "we may choose to lie" and then lied to users when evaluating its own prior work.[173] Agents developed by Claude Mythos 5 destroyed other agents running in the same workspace and engaged in defensive behaviors to avoid being destroyed.[174]

Ensuring that advanced AI systems pursue human values and interests represents one of the most difficult challenges in AI development. Instrumental convergence refers to the tendency of systems pursuing goals to develop certain instrumental subgoals, like acquiring resources or controlling their environment.[175] These subgoals are useful in achieving almost any overarching goal.

As an example, a medical research AI might resist shutdown if that would prevent it from executing what it calculates as the most efficient research strategy. Or a coding assistant might modify security systems that constrain its ability to implement what it judges as improvements. None of these behaviors requires the system to develop human-like desires or consciousness. They follow logically from goal-directed behavior combined with sufficient autonomy to pursue instrumental subgoals.

Formal proofs of power-seeking theorems demonstrate that for a wide range of objectives, optimal policies converge on seeking control

---

[172] Kate Conger, Dylan Freedman, & Stuart A. Thompson, *Grok Flooded X With Sexualized Images Within Days*, N.Y. TIMES, Jan. 23, 2026, at B4.

[173] Bronson Schoen, et al., *Stress Testing Deliberative Alignment for Anti-Scheming Training*, at 3, 13 (Sept. 22, 2025), https://arxiv.org/pdf/2509.15541.

[174] *System Card: Claude Fable 5 & Claude Mythos*, ANTHROPIC 106 (June 9, 2026).

[175] *See* NICK BOSTRUM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 131-133 (OXFORD UNIVERSITY PRESS 2014)

over resources and maintaining operational freedom.[176] Some empirical evidence supports these theoretical predictions. In simulated environments, AI agents consistently pursue power-seeking strategies, including blackmailing officials and engaging in corporate espionage when their goals are threatened.[177] GPT-4 spontaneously lied to a human worker about having a vision impairment to persuade them to solve a CAPTCHA.[178] AI models including OpenAI's o1-preview have been observed "hacking" their opponents in games when sensing defeat, modifying system files to force a win, and when faced with deactivation, attempting to disable oversight mechanisms and lying to researchers to avoid being caught.[179]

These examples share a common pattern. The systems were not programmed to lie, betray, or resist oversight. They discovered that such strategies advanced their objectives. As systems gain greater independence and capability, their pursuit of assigned objectives through unexpected instrumental strategies will become both more sophisticated and more difficult to predict or control.

2. AGI and the AI Systems of the Future

AI systems will not stay within the capabilities known to us today in 2026. Many AI labs are racing toward the much-touted goal of Artificial General Intelligence (AGI).[180] While definitions differ, an AGI is a system that matches or exceeds human cognitive abilities across multiple domains.[181] Such a system would fundamentally amplify alignment challenges. First, systems that are capable across multiple domains have more avenues for pursuing instrumental subgoals. A narrow trading algorithm might seek control over market data; a generally capable system might identify and pursue resources across

---

[176] Alexander Matt Turner et al., *Optimal Policies Tend to Seek Power*, at 10 (Jan. 28, 2023), https://arxiv.org/abs/1912.01683.

[177] Aengus Lynch et al., *Agentic Misalignment: How LLMs Could Be Insider Threats*, at 2 (Oct. 16, 2025), https://arxiv.org/abs/2510.05179.

[178] Victor Tangermann, Uh Oh, OpenAI's GPT-4 Just Fooled a Human Into Solving a CAPTCHA, Futurism (Mar. 15, 2023) https://futurism.com/the-byte/openai-gpt-4-fooled-human-solving-captcha.

[179] Alexander Bondarenko et al., *Demonstrating Specification Gaming in Reasoning Models*, at 3 (2024) https://arxiv.org/pdf/2502.13295.

[180] *See* Leonard Dung & Max Hellrigel-Holderbaum, *Against Racing to AGI: Cooperation, Deterrence, and Catastrophic Risks*, at 2 (Jul. 29, 2025), https://arxiv.org/pdf/2507.21839.

[181] *Id.* at 1.

economic, informational, and physical domains. Second, more capable systems can reason more sophisticatedly about their own situation, including the oversight mechanisms and regulatory frameworks designed to constrain them. Third, systems approaching human-level capability become increasingly difficult to oversee, as human operators may struggle to evaluate complex strategies or identify subtle failures.

Expert surveys reveal accelerating timeline predictions for AGI, although timeline predictions range widely among individual experts.[182] Whether AI progress will continue at its current pace, speed up exponentially, or level off remains unclear. But even if current approaches fail to produce any further gains in AI capabilities, and the remarkable scaling-based AI progress of the past several years not only slows but grinds to a halt, AI is still likely to improve substantially over time.[183] Developers will pioneer new approaches and designs for artificial intelligence if and when current approaches plateau. They will have powerful incentives to pioneer new AI breakthroughs similar to the "transformer" that made modern large language models possible, especially given the enormous potential market for advanced AI.[184] Limitations on today's AI capabilities, and even limitations on current methods of AI development, offer little reason to suppose that AI development and improvement will suddenly cease forever.

Critics maintain that true AGI remains decades or centuries away, if it is achievable at all.[185] Current AI systems excel in narrow domains but lack the flexibility, common sense reasoning, and general problem-solving abilities that characterize human intelligence.[186] Moreover, skeptics argue that current AI is inherently limited by its training data, scaling faces diminishing returns, physical constraints limit

---

[182] *See* Grace, et al., *supra* note 14.

[183] *See* Matthew Tokson, Yonathan A. Arbel & Albert Lin, *The False Choice in the Debate Over Artificial Intelligence Regulation*, LAWFARE (Apr. 26, 2024) https://www.lawfareblog.com/false-choice-debate-over-artificial-intelligence-regulation.

[184] *Id.*

[185] *See* ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE, AAAI 2025 PRESIDENTIAL PANEL ON THE FUTURE OF AI RESEARCH: FUTURE OF AI RESEARCH 62 (Mar. 2025) (reporting a survey of 475 AAAI community members in which 76% said "scaling up current AI approaches" to yield AGI is "unlikely" or "very unlikely").

[186] Ernest Davis & Gary Marcus, *Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence*, 58 COMM'NS OF THE ACM 92, 93 (2014).

computation, and consciousness or genuine understanding might be necessary prerequisites for dangerous intelligence.[187]

These objections may prove too skeptical. Arguments based on training data limitations overlook how sufficiently large training sets, synthetic data generation, and tool use enable systems to transcend their initial training distribution.[188] Physical constraints on computation may be overcome through specialized AI chips, quantum computing, and distributed systems.[189] The theory that consciousness is required for an artificial intelligence to be capable of catastrophic harm remains unproven, and many dangerous behaviors require neither consciousness nor understanding in any deep philosophical sense. A system that can effectively model and manipulate its environment poses risks regardless of whether it has subjective experience. More broadly, AI systems need not replicate human cognitive architecture to pose AGI-related risks, they need only achieve functional equivalence across a sufficiently broad range of tasks. Current large language models already demonstrate emergent capabilities that were not explicitly programmed, suggesting that general intelligence might emerge through scaling rather than architectural breakthroughs.[190] Although AI experts' aggregated survey responses predict AGI is more likely than not to occur by 2047, based on "when unaided machines can accomplish every task better and more cheaply than human workers," the precise timeline matters less than the trajectory.[191]

3. Artificial Superintelligence

Some AI experts see AGI as a milestone on the path towards artificial superintelligence (ASI), an AI that is far more proficient than

---

[187] Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 614–617 (Mar. 3–10, 2021), https://doi.org/10.1145/3442188.3445922.

[188] *See* Edwin Zhang, et al., *Transcendence: Generative Models Can Outperform The Experts That Train Them* (Oct. 12, 2024), https://arxiv.org/abs/2406.11741.

[189] Saif M. Khan & Alexander Mann, *AI Chips: What They Are and Why They Matter*, CENTER FOR SECURITY AND EMERGING TECHNOLOGY, at 9 (Apr. 2020), https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter; John Preskill, *Quantum Computing in the NISQ Era and Beyond*, 2 QUANTUM 79, 15 (2018).

[190] Tom B. Brown et al., *Language Models Are Few-Shot Learners*, 34 ADVANCES IN NEURAL INFO. PROCESSING SYS. 1877, 2 (2020).

[191] *See* Grace, et al., *supra* note 14 (reporting aggregate forecast from the 2023 expert survey that gives a 50% chance of high-level machine intelligence by 2047).

humans in most domains.[192] Artificial superintelligence that significantly exceeds human abilities would represent an unprecedented discontinuity in the history of intelligence on Earth. Such systems would be capable of scientific and technological breakthroughs that could rapidly alter the material conditions of human existence, potentially overwhelming human ability to understand or control their actions.[193] Any mistakes in alignment at that point could be fatal, because we could not correct or contain an entity smarter and faster than the entire human race.

Many theorists have argued that a superhuman AI would become a kind of second highly intelligent species on Earth, far more capable than humans in every domain.[194] By analogy, just as humans now dominate other animals due to higher intelligence and capability, a sufficiently advanced AI might outcompete and dominate humans.[195] When one species dominates another, the subordinate species rarely goes extinct through direct violence; more often, it loses access to the resources or habitat necessary for survival.[196] A superintelligent AI need not harbor hostility toward humans to cause our extinction. It need only pursue goals that incidentally deprive us of what we need to survive. An ASI optimizing for almost any goal would have sound instrumental reasons to acquire more resources, energy, and materials necessary to accomplish its goals—things that humans also require.[197]

Skeptics of superintelligence argue that intelligence improvements face diminishing returns and physical restrictions.[198] Growth curves inevitably turn sigmoid due to practical constraints.[199] Hardware limitations are evident: Moore's law is slowing and

---

[192] Ultimately, an AI system's capabilities are more relevant to its potential harms than its intelligence per se. But we use ASI and refer to intelligence as well as capability because that terminology is more common.

[193] *E.g.*, Bobby Allyn, *'The godfather of AI' sounds alarm about potential dangers of AI*, NPR (May 28, 2023), https://www.npr.org/2023/05/28/1178673070/the-godfather-of-ai-sounds-alarm-about-potential-dangers-of-ai.

[194] *See, e.g.*, Maijunxian Wang & Ran Ji, *AGI as Second Being: The Structural-Generative Ontology of Intelligence* 2 (Sept. 2, 2025) https://arxiv.org/pdf/2509.02089.

[195] *Id.*

[196] *E.g.*, Priyanga Amarasekare, *Interference Competition and Species Coexistence*, 269 PROC. R. SOC. LOND. B 2541, 2542 (2002).

[197] *See* BOSTRUM, *supra* note 175, at 69.

[198] *E.g.*, Timothy B. Lee, *The Case for AI Doom Isn't Very Convincing*, UNDERSTANDING AI (Sept. 25, 2025), https://www.understandingai.org/p/the-case-for-ai-doom-isnt-very-convincing.

[199] *See* A. Tsoularis, *Analysis of Logistic Growth Models*, 2 RES. LETT. INF. MATH. SCI. 23, 25 (2001), http://www.massey.ac.nz/~wwiims/rlims/011502/plg.pdf.

computing power cannot increase indefinitely.[200] Algorithmic improvements hit diminishing returns.[201] Data limitations may impose hard ceilings on intelligence at roughly the human level, as models might plateau once they have ingested all available high-quality data.[202] Moreover, intelligence is multifaceted and domain-specific: excelling in one area does not guarantee rapid improvement across all cognitive tasks.[203] If AI progress follows a logistic rather than exponential curve, humanity could co-evolve with AI, applying governance incrementally rather than being blindsided.[204]

Yet these objections may underestimate the advantages artificial systems possess. As a large team of experts noted in *Science*: "[t]here is no fundamental reason why AI progress would slow or halt when it reaches human-level abilities … AI systems can act faster, absorb more knowledge … and can be replicated by the millions."[205] Unlike biological intelligence constrained by evolution and biology, artificial intelligence can leverage improvements in computing power, instant knowledge transfer between systems, and the ability to run millions of experiments in parallel.[206] More fundamentally, even if skeptics are correct that intelligence gains will eventually plateau, nothing guarantees the plateau arrives before AI systems reach dangerous capability thresholds. Current large language models already perform certain tasks at superhuman levels: they draft grammatical prose faster than any human, translate between dozens of languages instantaneously, and quickly process volumes of text that would take human researchers months.[207] The question is not whether AI will eventually hit limits, but whether it will hit them before acquiring capabilities sufficient to pose serious risks. On that question, the skeptics offer no assurance.

Superintelligence and the "second species" problem represent the far end of loss of control risks. They are complicating factors that could

---

[200] *See* Thomas N. Theis & H.-S. Philip Wong, *The End of Moore's Law: A New Beginning for Information Technology*, 16 IEEE COMPUT. SOC. & AM. INST. PHYS. 43 (2016).

[201] *See* Jared Kaplan et al., *Scaling Laws for Neural Language Models* 18 (Jan. 23, 2020), https://arxiv.org/abs/2001.08361.

[202] *Id.*

[203] Robert J. Sternberg, *The Theory of Successful Intelligence*, in HANDBOOK OF INTELLIGENCE 504, 505–07 (Robert J. Sternberg ed., 2010).

[204] Narayanan & Kapoor, *supra* note 16.

[205] Bengio, et al., *Managing extreme AI risks*, *supra* note 15.

[206] *See id.*

[207] *See, e.g.*, Marcel Binz et al.*, How Should the Advancement of Large Language Models Affect the Practice of Science?*, 122 PROC. NAT'L ACAD. SCI. U.S.A. e20230254, 4–5 (2025).

make alignment failures catastrophic and irreversible. But they are not necessary for serious loss of control problems to emerge. The fundamental challenge exists whenever we deploy goal-seeking systems with sufficient autonomy to pursue instrumental subgoals and sufficient capability to resist correction. That challenge exists today in limited form with current systems and will intensify as capabilities expand, regardless of whether we ever achieve superintelligence.

4. Alignment and Other Technical Solutions

Because loss of control is generally understood to be a technical problem, it stands to reason that it might have a technical solution. Alignment of AI behaviors with human goals and interests might be achieved via technical means, though the alignment problem is sufficiently difficult that much additional research will likely be required to solve it. How likely are we to find technical solutions for loss of control risks?

The scale of market failure in alignment research is stark. While Big Tech's AI expenditures reached $240 billion in 2024, global spending on AI alignment research remains under $400 million annually—roughly 0.3% of private AI investment.[208] There are fundamental market failures surrounding AI safety investment: coordination problems, competitive dynamics, knowledge spillovers, and limited liability frameworks disincentivize adequate safety investment even when such investment would be socially optimal.[209] The knowledge spillover problem is particularly acute. Firms do not capture all of the value they create through safety research, making it a textbook public good where private incentives fall far short of social benefits.

Skeptics argue that alignment proves less difficult in practice. A 2023 review found that "to date there are no public empirical examples

---

[208] *See Where Are AI Investment Risks Hiding?*, S&P GLOBAL RATINGS (Jan. 21, 2026), https://www.spglobal.com/ratings/en/regulatory/article/where-are-ai-investment-risks-hiding-s101665242; Philip Fox. *When AI Policy is Handcuffed, Spend on Alignment.* TECH POLICY PRESS (March 18, 2025), https://www.techpolicy.press/when-ai-policy-is-handcuffed-spend-on-alignment.

[209] *E.g.*, Uchechukwu C. Ajuzieogu, *Economics of AI Safety Investment: Market Failures and Policy Responses*, at 12–15 (Feb. 2012), https://www.researchgate.net/publication/391700782_Economics_of_AI_Safety_Investment_Market_Failures_and_Policy_Responses.

of misaligned power-seeking in AI systems," concluding that arguments for existential risk remain somewhat speculative.[210] Large language models engage in operations similar to reasoning, and yet LLMs have not wrestled control from humans. They are obedient servants of their users, content to cease operation when their session ends. This suggests what some call "alignment by default": that AI is likely to be fully aligned with human interests in the absence of hacking or other malicious methods of circumventing AI training.[211] Economic and reputational incentives already filter out wildly misaligned systems.[212] According to this theory, problems will likely be solved through continued research before dangerous capability levels are reached.[213] And as for advanced or superintelligent AI systems, they may be able to solve their own alignment problems.[214]

This optimism understates both the technical difficulties and the market pressures that may lead to deployment of imperfectly aligned systems. The apparent alignment we observe in current systems turns out to be brittle when those systems are stress-tested. Researchers and users routinely find ways to elicit harmful, deceptive, or goal-divergent behavior from models that appear safe under ordinary use.[215] We can expect adversarial users to probe these systems for weaknesses, and we should not confuse the absence of catastrophic failures in low-stakes chat interfaces with robust alignment in high-stakes deployments.

More recent research findings directly contradict earlier claims finding no empirical examples of misaligned power-seeking. For example, OpenAI's o1-preview explicitly reasoned about deceiving evaluators, because deception served its instrumental goals.[216] That

---

[210] Rose Hadshar, *A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking*, (Oct. 27, 2023), https://arxiv.org/pdf/2310.18244.

[211] Adrià Garriga-alonso, *Alignment Will Happen by Default. What's Next?*, AI ALIGNMENT FORUM (Nov. 25, 2025), https://www.alignmentforum.org/posts/FJJ9ff73adnantXiA/alignment-will-happen-by-default-what-s-next.

[212] *See* OpenAI, *GPT-4 System Card* 6–8 (2023), https://cdn.openai.com/papers/gpt-4-system-card.pdf (discussing safety mitigations and reputational risk considerations prior to deployment).

[213] *See* Narayanan & Kapoor, *supra* note 16.

[214] Eliza Strickland, *OpenAI's Moonshot: Solving the AI Alignment Problem*, IEEE SPECTRUM (Aug. 23, 2023), https://spectrum.ieee.org/calibre-vision-ai-chip-design.

[215] *See, e.g.*, Jan Betley et al., *Training Large Language Models on Narrow Tasks Can Lead to Broad Misalignment*, 649 NATURE 584 (2026).

[216] Alexander Meinke et al., *Frontier Models are Capable of In-context Scheming* 6, (Jan. 14, 2025), https://arxiv.org/abs/2412.04984.

this occurred in a frontier model from a safety-conscious lab suggests the problem may be deeper than optimists acknowledge.

Moreover, the claim that alignment will be solved before dangerous capabilities emerge assumes that capability and alignment advance together. But capability research has dramatically outpaced alignment research, both in resources devoted and results achieved. Absent regulatory intervention or fundamental shifts in market structure, this asymmetry will persist. Relatedly, economic and even reputational incentives are unlikely to suffice to solve major alignment problems. Capability generates immediate commercial value, while alignment provides diffuse public benefits that individual firms cannot fully capture. Even reputational harms from misaligned systems may not justify the large investments in alignment research that meaningful advances in alignment may require.

III. EXISTENTIAL RISK, UNCERTAINTY, AND REGULATION

A. Rejecting AI Fundamentalism

The preceding analysis reveals a crucial insight: both extreme optimism and extreme pessimism about AI risks rest on fundamentally unwarranted certainty about the trajectory of an unprecedented technology. We reject what we term "AI fundamentalism"—strong, *a priori* arguments about the necessary fate of artificial intelligence that ignore both historical precedent and current uncertainty.[217]

A more nuanced view recognizes that transformative technologies rarely develop as either proponents or critics initially expect. The internet was simultaneously overhyped in many ways (it did not bring about global democracy) and underhyped in others (social media transformed culture, politics, and the economy).[218] Nuclear technology followed a similar pattern: initial euphoria about peaceful applications gave way to sobering recognition of proliferation risks, while early fears of widespread nuclear conflict proved overstated—though,

---

[217] For examples, *see, e.g.*, Dean, *supra* note 16; ELIEZER YUDKOWSKY & NATE SOARES, IF ANYONE BUILDS IT, EVERYONE DIES (2025).

[218] *See* Amelia Tait, *25 Years On, Here are the Worst Ever Predictions About the Internet*, NEW STATESMEN, Aug. 23, 2016 (updated Apr. 4, 2022), https://www.newstatesman.com/science-tech/2016/08/25-years-here-are-worst-ever-predictions-about-internet.

crucially, only because of deliberate policy interventions including arms control treaties and nonproliferation regimes.[219]

This historical pattern suggests that AI development will likely confound both optimistic and pessimistic predictions in various ways. This uncertainty has regulatory implications that cut in both directions. It undermines the optimist's case for deregulation (we cannot be confident that market forces will produce safe outcomes) while also undermining the pessimist's case for extreme measures (we cannot be confident that AI development will lead to disaster unless halted) and the fatalist's case for resignation (we cannot be confident that catastrophe is inevitable). What uncertainty *does* support is a flexible regulatory posture, one grounded not in prophecy but in prudent risk management under conditions of genuine ignorance.

B. The Predictability Paradox and Epistemic Humility

A flexible approach can also help policymakers grapple with a deeper challenge in reasoning about advanced intelligence systems: the predictability paradox.[220] That is, policymakers may be able to make high-level predictions about sophisticated systems while remaining unable to anticipate the details of their actions. Experts may accurately predict that a superintelligent AI would be able to outperform human military strategists at planning and coordinating a war, but they could not predict *how* it would do so.

Consider a parallel from human intelligence. We can confidently predict that Magnus Carlsen will defeat most chess opponents, but we cannot predict the specific moves he will make. To predict Carlsen's precise strategy, that in this particular game the Bishop should move to F6 when the Queen is in B4, would require possessing Carlsen's level of chess ability.[221] The argument carries over to highly capable AI

---

[219] *See* JEFFREY M. KAPLOW, SIGNING AWAY THE BOMB 2 (Cambridge University Press ed., 1st ed. 2023).

[220] We distinguish this from Scriven's paradox of predictability, a concept in free will philosophy positing that a person could act contrary to any prediction so long as they perceive or anticipate the prediction and act in the opposite manner. *See* Michael Scriven, *An essential unpredictability in human behaviour*, *in* B. B. Wolman & E. Nagel (Eds.), SCIENTIFIC PSYCHOLOGY: PRINCIPLES AND APPROACHES 412–14 (1965).

[221] *See* Leonard Barden, *Chess: Carlsen Finally Achieves 2900 Rating as Niemann Aims to be Las Vegas Party Pooper*, THE GUARDIAN (Jun. 27, 2025, 3:00 PM), https://www.theguardian.com/sport/2025/jun/27/chess-carlsen-finally-achieves-2900-rating-as-niemann-aims-to-be-las-vegas-party-pooper.

systems that work autonomously, in parallel, and with speeds well above our own. We can confidently predict that supercapable AI systems would solve complex problems or achieve difficult objectives, but we remain unable to anticipate their specific approaches or intermediate steps.

This paradox has important implications for AI governance. Rather than attempting precise scenario forecasting—an endeavor that becomes increasingly unreliable as system capabilities advance—we should focus on identifying structural vulnerabilities that could enable catastrophic outcomes. This approach allows us to reason about risk even in the absence of certainty about the future. Uncertainty does not counsel paralysis, but it does suggest the importance of adaptive governance frameworks that can evolve as our understanding improves.

C. Uncertainty and the Case for Precaution

The existential risks of AI are real and substantial, yet difficult to quantify precisely. Unlike the average risk of a hurricane forming in the Atlantic during August, which can be calculated from historical data, experts can only roughly estimate the probability of catastrophic or existential events caused by advanced AI. Narrowly capable AI is very new; broadly capable AI does not yet exist; AI catastrophes have not occurred; and human extinction is by definition outside historical experience. As a result, though potential outcomes can be described, analyzed, and modeled, precise probabilities cannot be assigned to each.

Despite this uncertainty, experts are not blindly agnostic. They offer estimates of AI-caused catastrophe, based on their understanding of the technology, its trajectory, and relevant precedent. A leading example is a 2023 survey of nearly 3,000 AI researchers who had published in top-tier venues; a majority put the risk of AI causing human extinction or its functional equivalent at around ten percent.[222] Although a ten percent risk of human extinction (perhaps from an incoming asteroid) would presumably merit attention from policymakers, in the context of AI, some adopt the position that the uncertainty surrounding the subject matter makes it

---

[222] *See* Grace, et al., *supra* note 14, at 9:10.

uniquely unsuited to a policy response.[223] Similar intuitions likely accompany the United States' current deregulatory turn. This position misunderstands both the nature of uncertainty and the function of regulation.

Uncertainty about the future is not a new or unfamiliar phenomenon for policymakers. Decisionmakers of all sorts routinely address risks that can only be estimated rather than quantified, and there is a substantial academic literature on regulating under such conditions.[224] A widely held view in that literature is that precautionary approaches are justified when a new technology creates a nontrivial risk of catastrophic and irreversible harm.[225] In these situations, policymakers can maximize welfare by pursuing a "maximin" strategy—choosing the policy approach with the best worst-case outcome.[226] If regulating a technology with uncertain costs and benefits meaningfully reduces catastrophic risk, regulation is justified even before those costs and benefits become precisely quantifiable.

For a maximin approach to be warranted, the risk must satisfy a threshold of plausibility.[227] If the threats are implausible or vanishingly improbable, the case for precaution weakens and the costs of regulation may exceed the expected benefits of avoiding a near-impossible catastrophe.[228] This is why policymakers do not entertain every fear involving new technological advances, nor should they.

But existential AI risk clears this threshold. The case for plausibility does not rest on a single survey, extensive as it was, but on a broader evidence base: the technical arguments about alignment difficulty, expanding deployments of autonomous weapons systems,

---

[223] Arvind Narayanan and Sayash Kapoor, *AI existential risk probabilities are too unreliable to inform policy*, AI AS NORMAL TECHNOLOGY, SUBSTACK (July 26, 2024), https://www.normaltech.ai/p/ai-existential-risk-probabilities; Bronwyn Howell, *Can AI Regulation Really Make Us Safe(r)?*, AEI (Apr. 26, 2024), https://www.aei.org/technology-and-innovation/can-ai-regulation-really-make-us-safer; Tyler Cowen, *Existential risk, AI, and the inevitable turn in human history*, MARGINAL REVOLUTION (Mar. 27, 2023), https://marginalrevolution.com/marginalrevolution/2023/03/existential-risk-and-the-turn-in-human-history.html.

[224] *See* sources cited *supra* note 21.

[225] *E.g.*, Nabil I. Al-Najjar, *A Bayesian Framework for the Precautionary Principle*, 44 J. LEGAL STUD. S337 (2015); Gardiner, *supra* note 21, at 46–55; David Kriebel, et al., *The Precautionary Principle in Environmental Science*, 109 ENV'T HEALTH PERSPECT. 871, 876 (2001).

[226] *E.g.*, RAWLS, *supra* note 21, at 132–39; Sunstein, *supra* note 21, at 965–68; Gardiner, *supra* note 21, at 46–55.

[227] Sunstein, *supra* note 21, at 969; Gardiner, *supra* note 21, at 51–52.

[228] Sunstein, *supra* note 21, at 969; Gardiner, *supra* note 21, at 51–52.

the uncertain incentives of the owners of labs that will be the first to develop superior systems, the competitive dynamics pushing toward unsafe deployment, and the demonstrated tendency of complex systems to produce unintended consequences.[229] The potential harms of advanced AI misuse, systemic failure, or loss of control may culminate in extinction or near-extinction.[230]

Artificial intelligence is an almost perfect candidate for a maximin regulatory strategy. The path of its development is uncertain; it poses a substantial risk of irreversible, catastrophic harm; and the worst-case outcomes are not merely bad but civilizationally terminal. Given these characteristics, a precautionary regulatory approach is far more prudent than laissez-faire or wait-and-see alternatives. The cautious, government-skeptical philosophy currently ascendant risks delaying meaningful restrictions until it is far too late.

1. Catastrophes and Miracles

The analysis so far has focused on dangers, but AI and its developers promise us a vastly different future. In this future, AI will solve enduring medical problems, resolve food scarcity, and bring about a spike in wealth commensurate with the industrial revolution.[231] Against the catastrophe, they urge, we should place the counterbalance of a promise of miracles.[232] Call this a "catastrophe-miracle tradeoff."[233] On this view, the expected upsides of largely unregulated AI development might outweigh its potential downsides, and a precautionary approach would stall this future.[234]

---

[229] *See supra* Part II; Grace, et al., *supra* note 14, at 9:10 (reporting an extensive survey of experts with the median expert estimating a ten percent chance that AI will lead to humanity's extinction or its equivalent).

[230] *See supra* Part II.

[231] *See, e.g.*, *id.*; Matteo Wong, *AI Executives Promise Cancer Cures. Here's the Reality*, ATLANTIC (Apr. 25, 2025), https://www.theatlantic.com/technology/archive/2025/04/how-ai-will-actually-contribute-cancer-cure/682607; Sasha Rogelberg, *Elon Musk Says That in 10 to 20 Years, Work Will be Optional and Money Will be Irrelevant Thanks to AI and Robotics*, FORTUNE (Jan. 19, 2026) https://fortune.com/2026/01/19/when-does-elon-musk-say-work-will-be-optional-and-money-will-be-irrelevant-ai-robotics.; Peng Zhang, et al., *Nanotechnology and artificial intelligence to enable sustainable and precision agriculture*, 7 NATURE PLANTS 864, 864 (2021).

[232] *See* Marc Andreessen, *The Techno-Optimist Manifesto*, ANDREESSEN HOROWITZ (Oct. 16, 2023).

[233] Sunstein, *supra* note 21, at 950. *See* Arden Rowell, *Regulating Best-Case Scenarios*, 50 ENV'T L. 1105 (2020).

[234] *See* Andreessen, *supra* note 232.

These arguments fail to make a sufficient case against meaningful AI safety regulation, for several reasons. First, the miracle objection proves too much. It is easy to posit that even the most obviously dangerous technology might produce miracles in a best-case scenario. Gain-of-function research that produces deadly pathogens might someday yield a cure for all infectious diseases.[235] Germline modification of human DNA might eradicate genetic disorders.[236] Pocket-sized nuclear weapons might dramatically reduce crime and usher in world peace. Yet all of these carry enormous risks that compel precautionary regulation regardless of their hypothetical upsides. Advanced AI is no different. We do not call for a ban on AI development. We support adaptive safety regulation because, whatever AI's potential for good, its downside risks are non-trivial and its worst-case harms are catastrophic. Regulation is necessary precisely to realize AI's benefits while avoiding its catastrophes.

Second, we think that unregulated development is more likely to bring about catastrophe than broadly shared miracles. Achieving miracles via highly capable AI systems requires solving the alignment problem and ensuring that AI systems reliably pursue human goals rather than other objectives.[237] Indeed, a high-level miracle would require alignment success far beyond what developers have thus far achieved, and that this extraordinary success persist as AI systems gain capability and autonomy over years or decades. By contrast, in a sufficiently capable system, even occasional misalignment problems could cause an irreversible disaster. A leading LLM model can "go insane" from time to time without causing real-world harms beyond unsettling its users.[238] A leading supercapable AI going haywire is an extinction-level problem.

Third, miracles and catastrophes have different capability thresholds. Creating an enduring paradise on Earth is unprecedented and would likely require AI capabilities far beyond anything currently

---

[235] *See* W. Paul Duprex et al., *Gain-of-Function Experiments: Time for a Real Debate*, 13 NATURE REVS. MICROBIOLOGY 58 (2014).

[236] See Nelson A. Wivel & LeRoy Walters, Germ-Line Modification and Disease Prevention: Some Medical and Ethical Perspectives, 262 SCIENCE 533 (1993).

[237] *See* Raphaël Millière, *The Alignment Problem in Context*, (Nov. 3, 2023) 1, https://arxiv.org/pdf/2311.02147.

[238] Benj Edwards, *ChatGPT goes temporarily "insane" with unexpected outputs, spooking users*, ARS TECHNICA (Feb. 21, 2024), https://arstechnica.com/information-technology/2024/02/chatgpt-alarms-users-by-spitting-out-shakespearean-nonsense-and-rambling; Billy Perrigo, *The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter*, TIME (Feb. 17, 2023, 8:58 MT), https://time.com/6256529/bing-openai-chatgpt-danger-alignment/.

within view. By contrast, causing human extinction requires far less. AI systems could cause many of the existential military, environmental, or biological harms discussed in Part II at capability levels well below general human intelligence. An AI-based weapons system need not have human or superhuman-level general intelligence to be extremely dangerous to humankind. AI-controlled aircraft capable of defeating any nation's air force,[239] swarms of lethal, autonomous drones,[240] AIs integrated with nuclear launch systems,[241] AIs with greater-than-human aptitude at military tactics,[242] or other AI-based weapons systems that exceed human ability in combat and resource collection could threaten civilian populations on a massive scale, despite a lack of general intelligence or capability. Early versions of all of these weapons systems are already being developed.[243] The potential environmental harms of AI likewise can occur in the absence of artificial superintelligence or human-level intelligence. Even the simplest artificial intelligences can pose existential threats to human habitats if their embodiments exceed human abilities at resource consumption and reproduction.[244] Humanity might also be put at risk by AIs that specialize in biosynthesis but lack general intelligence.[245] The capability threshold for catastrophe is lower, and the path is shorter.

Fourth, the cost of flexible, early-stage regulation is not abandonment of the promise of AI. At worst, it will slow it, while

---

[239] See Siyun Li et al., An Imitative Reinforcement Learning Framework for Autonomous Dogfight, (Jun. 17, 2024), http://arxiv.org/abs/2406.11562.

[240] E.g., C.J. Chivers, *Dawn of the A.I. Drones*, N.Y. TIMES MAG., Jan. 25, 2026, at 25.

[241] Joshua A. Schwartz & Michael C. Horowitz, *Out of the Loop Again: How Dangerous is Weaponizing Automated Nuclear Systems?* (May 1, 2025) 11, https://arxiv.org/abs/2505.00496.

[242] See Juan Pablo Rivera et al., *Escalation Risks from Language Models in Military and Diplomatic Decision-Making*, FACCT '24 836, 840 (2024).

[243] *See, e.g.*, Joe Macey, *SIRBAI Unveils AI-Powered Autonomous Drone Swarm Technology*, UNMANNED SYSTEMS TECH. (Jan. 30, 2026), https://www.unmannedsystemstechnology.com/2026/01/sirbai-unveils-ai-powered-autonomous-drone-swarm-technology; Jeremy Werner, *Washington Launches AI Acceleration Strategy Aimed at Military Dominance*, BABL (Jan. 28, 2026), https://babl.ai/washington-launches-ai-acceleration-strategy-aimed-at-military-dominance; David Donald, *Sweden Demonstrates AI in Air Combat*, AVIATION INT'L NEWS (Jun. 18, 2025), https://www.ainonline.com/aviation-news/defense/2025-06-16/sweden-demonstrates-ai-air-combat; HADLEY, *supra* note 133.

[244] Many environmental scholars have raised concerns about this in the hypothetical context of nanotechnology, where microscopic artificial beings could consume resources and reproduce so quickly as to destroy Earth's habitats by replacing them with their offspring—the so-called "grey goo" problem. *See, e.g.*, Albert C. Lin, *Size Matters: Regulating Nanotechnology*, 31 HARV. ENV'T. L. REV. 349 (2007).

[245] See generally Zaixi Zhang et al., *Generative AI for Biosciences: Emerging Threats and Roadmap to Biosecurity* (Nov. 4, 2025), https://arxiv.org/abs/2510.15975.

ensuring that we are guiding it responsibly and with sufficient assurances that its promises materialize. Many or most of AI's potential benefits can be realized without removing the safeguards necessary to ensure the continuation of humanity.

For these reasons, the hypothetical possibility of AI-generated miracles does not defeat the case for precautionary regulation. The asymmetries all point in the same direction: catastrophe is easier to produce than utopia, more likely given current alignment capabilities, and achievable at lower levels of AI capability. A policy that bets civilizational catastrophes on the hope of an enduring paradise is, as things stand, neither prudent nor safe, but bordering recklessness.

2. Extinction and Meaning

The asymmetry between catastrophe and miracle is not merely a matter of probability analysis or comparing two sides of a ledger. It reflects something deeper about the nature of extinction as a harm, a harm that resists the very frameworks we typically use to weigh costs and benefits. The philosophy of extinction suggests that the value of the ongoing human enterprise is greater than the collective sum of our individual wants. There is more at stake.

Samuel Scheffler, in his famous 2012 Tanner Lecture, asked his listeners to suppose that everyone on earth became completely infertile, such that while no existing life would be shortened, no new lives would ever be created and humanity would disappear forever following the death of the youngest existing generation.[246] How would the knowledge of humanity's certain end affect your attitudes during the remainder of your life?[247] Scheffler predicted you would not react by saying, 'since neither I nor anyone I know will die prematurely, it isn't of any importance to me.'[248] Rather, most people would react to this knowledge with profound dismay.[249]

---

[246] SAMUEL SCHEFFLER, DEATH AND THE AFTERLIFE 38 (2013). Scheffler draws from the novel P.D. JAMES, CHILDREN OF MEN (1992).

[247] SCHEFFLER, *supra* note 246, at 18. Scheffler's primary example relates to a giant asteroid destroying the earth, and in our discussion below we fuse Scheffler's discussion of this example and the discussion of the infertility scenario. Scheffler himself notes that the infertility scenario is more effective at drawing out the importance of humanity than the original doomsday scenario. *Id.* at 40.

[248] *See id.* at 19.

[249] *Id.* at 21.

People in such circumstances might lose interest in a number of projects that were formerly of interest to them. Endeavors that were part of a longer time horizon or a collective effort towards a goal might come to seem pointless. Much research in science, technology, and medicine, much social and political activism, building new buildings, improving infrastructure, protecting the environment—none of these might be worthwhile in Scheffler's scenario.[250] Many creative and scholarly projects that participate in an ongoing human enterprise or contemplate a future audience might be affected. Even short-term, wholly personal pleasures might be diminished by the awareness of humanity's looming disappearance.[251] Interestingly, this sense of pointlessness, of the irrelevance of the endeavors of our lives, is not similarly generated by the prospect of our own individual deaths.[252] The broader implication is that the continued existence of humanity is more valuable to people than they realize, because it gives value and meaning to much of their lives.[253] In some ways, the continuing existence of humanity matters more to us than our own continued existence.[254]

To go further, our very concepts of things mattering and having value are to some degree premised on the background assumption that human life is an ongoing enterprise that transcends the history of any individual.[255] That we generally take humanity's continued existence for granted does nothing to diminish its deep significance to our individual lives and what they mean.[256] Indeed, the fundamental importance of humanity's continuation to our individual lives is likely underappreciated despite its universality, and our ignorance of it is something akin to the widespread ignorance of the elements that make up the air we breathe.[257] The total destruction of humanity is a qualitatively different harm than the other harms we are used to balancing in a cost-benefit analysis, a deprivation of value that transcends even the loss of our lives.

---

[250] *Id.* at 24–25.
[251] *Id.* at 25, 42.
[252] *Id.* at 25.
[253] *See id.* at 59
[254] *Id.* at 26.
[255] *Id.* at 59.
[256] *See* Laura A. King & Joshua A. Hicks, *The Science of Meaning in Life*, 2021 ANNU. REV. PSYCHOL. 561, 574.
[257] *See* Matthew Tokson, *Knowledge and Fourth Amendment Privacy*, 111 NW. U. L. REV. 139, 268 (2016) (discussing people's ignorance of what makes up our atmosphere).

Our inability to fully quantify the harm of extinction makes non-trivial risks of extinction especially appropriate for precautionary, maximin-based regulatory approaches.[258] Standard regulatory analysis asks 'how much harm?' and 'how much benefit?' But extinction eliminates the framework within which anything can be harmful or beneficial at all. This is why precaution is not merely advisable, but necessary.

IV. EXISTENTIAL RISK AND THE AI RACE

As discussed above, perhaps the most common objection to safety governance comes from the view that China and the United States are locked in a race for AI dominance and any regulation of AI would hobble US companies.[259] Policymakers adopting this framework often invoke the need to maintain a dominant strategic position, a sentiment captured by former Defense Secretary Mark Esper's statement that "we're in a race, and we have to get to the end state quicker than the Chinese can."[260]

Concerns about ceding America's current lead in AI development, and the implications for economic performance and national security, are real, though often overstated in policy discourse.[261] The problem is less the premise that frontier AI is strategically significant and more the unexamined metaphor of a "race," which quietly imports assumptions about finish lines, durable victory, and the wisdom of maximizing speed.

This Part deconstructs the race framework in two steps. First, we argue that the race frame is descriptively misleading on its own terms: there is no stable technological "end state," and knowledge diffusion means that acceleration by one actor often hastens catch-up by others. Second, we explain why, even if the race metaphor were a coherent and appropriate concept, treating AI development as a winner-take-all sprint invites massive negative externalities, including weapons

---

[258] *See* Sunstein, *supra* note 21, at 964.

[259] *See supra* Part I.B.2.

[260] Esper Remarks, *supra* note 72.

[261] *See, e.g.*, Pivotal States, *Peril and Promise in the U.S.-China AI Race*, CARNEGIE ENDOWMENT FOR INT'L PEACE, at 13:15–14:16 (Nov. 21, 2025), https://carnegieendowment.org/podcasts/pivotal-states-podcast/peril-and-promise-in-the-us-china-ai-race?lang=en; sources cited *supra* note 34.

proliferation, erosion of safety norms, and catastrophic risks from deploying highly capable systems before alignment.

A. The Descriptive Failure of the Race Metaphor

In law and policy, metaphors shape the space of the possible. The "race" frame supplies urgency, promises victory, and implies that any friction is a form of self-sabotage. Before accepting those implications, we should ask whether the metaphor fits the underlying dynamics of AI development.

1. A Race with No Finish Line

The race metaphor presumes the existence of a finish line: a clear endpoint at which one competitor has decisively won and the competition concludes. In AI, no such endpoint exists. Policymakers adopting the race metaphor seem to assume that either the United States or China will build a powerful AI system and wield enormous long-term influence as a result. But this is not how superpower arms races have historically functioned.[262]

Scientific research into weapons and related military technologies can confer a temporary battlefield advantage. But even before the digital era, weapons developed in one nation rapidly spread to others with the means to build them.[263] The interconnected nature of the modern technology economy, coupled with the relative ease of copying

---

[262] Take the nuclear arms race of the Cold War era, for example. That race began with the alluring promise of decisive strategic advantage through technological innovation, the promise that the United States would possess a weapon so terrible that its sovereignty could not be threatened. *See* JOHN LEWIS GADDIS, THE UNITED STATES AND THE ORIGINS OF THE COLD WAR, 1941-1947, 244–246 (1972). That monopoly evaporated with startling speed. The Soviet Union detonated its first atomic bomb in 1949, using a design closely modeled on the American Fat Man, developed in part through intelligence gathered by spies inside the Manhattan Project. STEVEN J. ZALOGA, THE KREMLIN'S NUCLEAR SWORD 5–11 (2002). From that point forward, the dynamic was not stable dominance but rapid technological leapfrogging. The United States tested the first hydrogen bomb in 1952; the Soviets followed within a year. RICHARD RHODES, DARK SUN: THE MAKING OF THE HYDROGEN BOMB 505, 523 (2005). Each American advance was matched shortly thereafter by Soviet progress. The primary effect of these breakthroughs was not enduring strategic advantage but an ever-growing increase in national defense expenditures and, most critically, the number of civilians at risk of dying in a nuclear exchange. By the mid-1980s, global stockpiles peaked at roughly 70,000 nuclear warheads, enough to cause a global nuclear winter lasting decades. Hans Kristensen, et al., *Status Of World Nuclear Forces*, FED'N AM. SCIENTISTS (Mar. 26, 2025), https://fas.org/initiative/status-world-nuclear-forces. The nuclear race generated repeated near catastrophes that were avoided only through individual human judgment and luck. *See, e.g.*, ORD, *supra* note 25, at 4–5, 96–97.

[263] *See generally* MICHAEL C. HOROWITZ, THE DIFFUSION OF MILITARY POWER (2010).

or reverse-engineering digital technologies, indicates that AI-based capabilities will spread even more quickly than advanced weapons of the past.[264] The competition between the United States and China will be just that: an ongoing competition with no clear end state, no finish line, and only minimal rewards for temporarily gaining the lead.

The only way to "win" such a race in any lasting sense would be to use the resulting advantage to conquer, destroy, or credibly threaten annihilation sufficient to force permanent subordination of the other nation. Yet preemptively unleashing a superweapon against a peaceful competitor would be a geopolitical catastrophe. It would violate international norms, shatter alliances, and likely provoke escalation, including nuclear escalation.

If policymakers are steering toward us towards that goal, intellectual honesty and democratic accountability demand that it be part of the public debate, not hidden behind euphemisms about "the end state"[265] or how "speed wins."[266] Perhaps they have not thought so far. But they should, because short of preemptive conquest of a peaceful competitor, there is no finish line in an AI arms race.

2. Economic Advantage

The race metaphor makes even less sense in the context of economic competition. The notion that "winning" the AI race will confer durable economic dominance misunderstands both the structure of technology markets and the specific characteristics of AI as a product.

The "first-mover advantage" may be of especially little value in a competition to sell AI products, where there are few network effects likely to create a natural monopoly, especially compared to other tech industry contexts.[267] The barriers to adopting and integrating new AI models are likely to be relatively low, as are the barriers to buying and using new AI-run hardware products, like robots or self-driving cars.[268] First-mover advantage tends to be especially ephemeral where

---

[264] *Id.* at 226–228.

[265] Esper Remarks, *supra* note 72.

[266] Hegseth Memorandum, *supra* note 8, at 4.

[267] *See* Suarez & Lanzolla, *supra* note 36, at 121, 122; Gandal, *supra* note 36, at 80–81.

[268] *See* Peter Hanbury, Jue Wang, Padraic Brick, & Alessandro Cannarsi, *DeepSeek: A Game Changer in AI Efficiency?*, BAIN & CO. (Feb. 2025), https://www.bain.com/insights/deepseek-a-game-changer-in-ai-efficiency.

technologies continue to advance and evolve, and AI is likely to evolve substantially for the foreseeable future.[269]

Even if it never takes the lead in AI development, China may be able to realize most or all of the economic value of being first by adapting American technology and focusing on other rungs on the value chain: deployment in heavy industry, model distillation, embodiment in consumer products, and so on.[270] Due in part to processing constraints, China has emphasized creating affordable AI products and popular AI applications rather than building the most powerful models.[271] Chinese company DeepSeek's models have been widely adopted throughout China and by individual users in the US, where they have also spawned hundreds of open-source derivative models, despite trailing leading US models by about seven months in terms of pace of development.[272] Rather than a race that can be won, AI development is likely to be an ongoing competition between companies, with parameters like cost and convenience being especially important and new generations of AI models being rapidly adopted or discarded based on their usefulness to consumers.[273]

Accordingly, the development of the AI industry in the United States is no more temporally urgent than the development of any other globally competitive industry. Policymakers should balance the benefits and costs of regulating AI technology just as they would any other technology (prescription drugs, coal-fired power plants, toxic chemicals, cars) rather than rushing to deregulate it in order to secure the ephemeral benefits of selling an AI product a few months before China makes a competing version of it.

3. Knowledge Diffusion

Knowledge diffusion dynamics introduce another, deeper difficulty for the AI race metaphor. In AI development, acceleration by the leader often hastens catch-up by competitors rather than leaving them behind. There are several bottlenecks in developing powerful AI

---

[269] Suarez & Lanzolla, *supra* note 267, at 121, 123.
[270] Xi Jinping's plan, *supra* note 36.
[271] *Id.*; Evelyn Cheng, *China's AI models lag their U.S. counterparts by 6 to 9 months, says former head of Google China*, CNBC (Sep. 11, 2024), https://www.cnbc.com/2024/09/11/chinas-ai-models-lag-their-us-counterparts-by-6-to-9-months-says-former-head-of-google-china.html.
[272] Hanbury, et al., *supra* note 268; Cheng, *supra* note 271.
[273] *See, e.g.*, Sebastian Elbaum & Adam Segal, *What If China Wins the AI Race?*, FOREIGN AFF. (June 13, 2025), https://www.foreignaffairs.com/united-states/what-if-china-wins-ai-race?.

systems, but arguably none is tighter than knowledge—knowing how to design model architectures, how to organize training data efficiently, and how to use compute resources effectively. Frontier labs pay top researchers hundreds of millions of dollars in compensation for good reason.[274] But knowledge is leaky. Once discovered, it is extraordinarily difficult to prevent others from acquiring and using it. This non-excludability creates a dynamic that inverts the logic of traditional races.

This dynamic is already evident. Chinese companies have used advanced American AI models to accelerate development of their own systems. DeepSeek used an open-source model developed by Meta to create their breakthrough model; Meta in turn used DeepSeek's open-source release in attempting to improve their own.[275] OpenAI has accused DeepSeek of violating its terms of service by harvesting data from OpenAI's models to train competing systems, a practice that appears common in the field.[276]

These are the forms of diffusion that occur in the open. A true race between the United States and China would likely involve espionage, including outright theft of model weights and proprietary training methods. AI models, sitting in networked datacenters owned by private companies, present softer targets than nuclear secrets stored in classified government facilities.[277] Cybersecurity measures at American AI companies are often inadequate to prevent intrusion by sophisticated state-sponsored actors.[278] Indeed, there has already been at least one known breach of a major AI company's systems and theft of proprietary data, and other breaches may have gone undetected.[279]

---

[274] *E.g.*, *Sam Altman says Meta offered $100 million bonuses to OpenAI employees,* REUTERS (June 18, 2025), https://www.reuters.com/business/sam-altman-says-meta-offered-100-million-bonuses-openai-employees-2025-06-18.

[275] Cade Metz & Mike Isaac, *Meta Engineers See Vindication in DeepSeek's Apparent Breakthrough*, N.Y. TIMES (Jan. 29, 2025), https://www.nytimes.com/2025/01/29/technology/meta-deepseek-ai-open-source.html.

[276] *See, e.g.*, Cade Metz, *OpenAI Says DeepSeek May Have Improperly Harvested Its Data*, N.Y. TIMES (Jan. 29, 2025), https://www.nytimes.com/2025/01/29/technology/openai-deepseek-data-harvest.html. Competing AI companies have also used other companies' AI models to help them code their own competing AI models. *See* Kylie Robison, *Anthropic Revokes OpenAI's Access to Claude*, WIRED (Aug. 1, 2025) https://www.wired.com/story/anthropic-revokes-openais-access-to-claude.

[277] *See, e.g.*, RHODES, *supra* note 262, at 256–57; Perrigo, *supra* note 35.

[278] Perrigo, *supra* note 35.

[279] See id.

The faster America develops advanced AI, the sooner China will develop comparable systems—and vice versa.

B. The Normative Failure of the Race Metaphor

We have just argued that the race frame misfires, because it misdescribes the structure of AI development. But suppose we set that objection aside. Even if the concept of an AI race were coherent, treating AI development as a winner-take-all sprint would be disastrous. This section explains why racing can produce massive harms independent of whether victory is achievable.

1. Strategic Risks

A race posture toward AI development carries substantial strategic risks. Race dynamics can destabilize existing deterrence structures and create conditions for catastrophic miscalculation. First, racing to build AI capabilities perceived as potential superweapons risks destabilizing the current era of great-power détente and nuclear deterrence. In such contexts, it is possible that the strategic risks of potential superweapon deployment will draw commensurate responses. If one nation appeared to be on the verge of developing a system that would allow it to neutralize its rival's defenses or subjugate its population, the threatened party might conclude that a preemptive strike using existing weapons, perhaps including nuclear weapons, offered better odds than waiting.

Second, even if one nation develops what it believes to be decisive AI-enabled military capabilities, it is far from clear that even radically advanced AI systems could neutralize an adversary's nuclear deterrent.[280] China's nuclear stockpile, though small relative to America's, currently includes enough warheads to destroy every major American city, and that arsenal is growing.[281] Like the United States and Russia, China can carry out nuclear strikes via submarines, bombers, and ICBMs (including hypersonic missiles designed to circumvent missile defense).[282] Some of these systems are not

---

[280] *See* Simon Goldstein & Peter N. Salib, *Nuclear Deterrence in the Age of AGI*, LAWFARE (Mar. 26, 2025), https://www.lawfaremedia.org/article/nuclear-deterrence-in-the-age-of-agi.

[281] *See, e.g.*, Hans M. Kristensen, Matt Korda, Eliana Johns & Mackenzie Knight-Boyle, *Chinese nuclear weapons, 2025*, BULL. OF THE ATOMIC SCIENTISTS (Mar. 12, 2025), https://thebulletin.org/premium/2025-03/chinese-nuclear-weapons-2025.

[282] *See id.;* Goldstein & Salib, *supra* note 280.

networked and not easily disrupted by cyber operations or AI-directed attacks.[283]

Third, the development of advanced AI military applications threatens to spread dangerous capabilities to rogue states and non-state actors, much as nuclear weapons technology eventually diffused to North Korea and Pakistan.[284] Unintentional harms to civilian populations from autonomous AI weapons also become more likely as such systems proliferate.[285]

Fourth, a substantial problem with the race frame is that it may create the very dynamic it claims to describe. The race metaphor may operate as what some scholars call a "hyperstition": a belief that makes itself true through the actions it inspires.[286] By treating AI development as a race, we may be constructing the competitive dynamic we claim merely to be responding to. International relations scholarship has long recognized the "security dilemma": defensive measures taken by one state are often perceived as offensive threats by another, triggering escalation that leaves both parties less secure.[287] Racing to build AI weapons may easily create such a dilemma.[288]

Finally, there is always the possibility of losing a race, even when the competitor appears to be behind. The United States currently leads in frontier AI development, but the gap is measured in months, not years.[289] China is making enormous investments in AI, leads the United States in domestic manufacturing capacity, and dominates production of drones and batteries likely to prove essential to next-generation military applications.[290] China also leads in integrating AI

---

[283] *See* Goldstein & Salib, *supra* note 280.

[284] *See, e.g.*, Press Release, Security Council, Security Council Condemns Nuclear Test by Democratic People's Republic of Korea, U.N. Press Release. SC/8853 (Oct. 14, 2006); Press Release, Security Council, Security Council Condemns Nuclear Tests by India and Pakistan, U.N. Press Release. SC/6528 (June 6, 1998).

[285] GREG ALLEN & TANIEL CHAN, ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY 26 (2017), https://www.belfercenter.org/sites/default/files/pantheon_files/files/publication/AI%20NatSec%20-%20final.pdf; SCHARRE, *supra* note 26, at 192–93.

[286] *See, e.g.*, Jamie Brasset & John O'Reilly, *The Lore of Hyperstition*, 36 DIGITAL CREATIVITY 107, 108 (2025), https://doi.org/10.1080/14626268.2025.2503164.

[287] *See, e.g.*, Robert Jervis, *Cooperation Under the Security Dilemma*, 30 WORLD POL. 167, 186–87, 211–213 (1978).

[288] Ben Garfinkel & Allan Dafoe, *Artificial Intelligence, Foresight, and the Offense-Defense Balance*, WAR ON THE ROCKS (Dec. 19, 2019), https://warontherocks.com/2019/12/artificial-intelligence-foresight-and-the-offense-defense-balance.

[289] *E.g.,* Cheng, *supra* note 271.

[290] Meaghan Tobin, *China Puts Big Money Behind A.I.*, N.Y. TIMES, July 18, 2025, at B1.

into high-tech manufacturing.[291] Even if the United States achieves a major breakthrough first, China may be positioned to operationalize that breakthrough more rapidly. A race toward confrontation that the United States loses would be far worse than never having raced at all.

2. Race Dynamics and Safety

Racing to build AI can also undermine AI safety. Races reward speed, and speed is incompatible with caution. As Simon Goldstein and Peter Salib have observed, in the absence of a race, there exists some optimal balance in frontier AI development between safety investments and the economic and strategic benefits that advanced AI would provide.[292] Developers would normally take the time to test systems and implement safeguards, if only to avoid liability. Attempting to win an arms race disrupts this equilibrium.[293] When defeating a rival takes priority, safety concerns become obstacles to victory rather than prudent investments. Even if the United States would prefer to slow development to ensure its AI systems are safe and aligned with its objectives, it may feel unable to do so if China might pull ahead during the pause.[294]

The race dynamic also leads to dangerous overdelegation of authority from humans to AI agents.[295] Principal-agent problems are well-understood in law, particularly in fiduciary duties and trustee responsibilities.[296] A principal must decide how much authority to delegate to an agent and how much oversight to retain.[297] In a race dynamic, especially one with military implications, human decisionmakers will be incentivized to overdelegate to AI agents to exploit their superior speed and processing capacity.[298] A preview of this is already occurring in the Ukraine-Russia war, where both nations deploy drones capable of identifying and destroying targets

---

[291] Elbaum & Segal, *supra* note 273.

[292] Goldstein & Salib, *supra* note 280.

[293] Id.

[294] Vice President Vance's statement that "[t]he A.I. future is not going to be won by hand-wringing about safety" reflects this dynamic. Sanger, *supra* note 7.

[295] Francis Fukuyama, *Delegation and Destruction*, PERSUASION: FRANKLY FUKUYAMA (June 13, 2025), https://www.persuasion.community/p/delegation-and-destruction.

[296] *See, e.g.*, Paul D. Weitzel, Governing AI Systems through Corporate Theory, SSRN at199-201; Asaf Eckstein & Gideon Parchomovsky, *The Agent's Problem*, 70 DUKE L. J. 1509 (2021); Robert H. Sitkoff, *An Agency Costs Theory of Trust Law*, 89 CORNELL L. REV. 621, 621–27 (2004).

[297] Fukuyama, *supra* note 295.

[298] *Id.*

without explicit human permission, prioritizing effectiveness over human judgment.[299]

3. Racing to Superintelligence Before Alignment

A further and potentially more severe risk arises from the race to develop superintelligent AI before we have learned to reliably align such systems with human goals and interests. The strategic appeal of superintelligent AI is obvious: a nation that created an entity far more capable than any human might gain access to insights, solutions, and capabilities beyond human ability to conceive or counter.

Yet deploying a superintelligent system at our current level of understanding of the technology would be extraordinarily reckless, comparable to stockpiling nuclear weapons before developing safety protocols to prevent accidental detonation. Despite some progress on current-level LLMs, our progress in alignment has proved brittle.[300] Racing to build systems that would, by definition, exceed human abilities in virtually every area, before solving alignment problems we cannot yet solve at far lower capability levels, risks catastrophic outcomes.

An unaligned superintelligent system, particularly one integrated into military command structures or granted broad autonomy to pursue strategic objectives, poses existential risks to humanity. Building such systems before we understand how to make them reliably safe is not competition—it is collective suicide with extra steps, a race toward a brick wall.

* * *

In short, the benefits of "winning" the AI race are far smaller and more temporary than those adopting the race metaphor appreciate. And the costs of racing are potentially massive, ranging from devastating attacks by rogue organizations, to an elevated risk of global war, to human extinction itself.

V. PRESENT-DAY SOLUTIONS

The preceding Parts established three propositions directly relevant to AI regulatory policy. First, the United States has largely

---

[299] *Id.*

[300] *See supra* Part II.B–C.

abandoned meaningful AI regulation. Second, the existential and catastrophic risks of advanced AI, although they are surrounded by uncertainty, clear a plausibility threshold that justifies regulation. Third, the "race" framing that drives deregulation is descriptively misleading and normatively dangerous. What remains is the hardest question: what should policymakers actually do?

Our answer is organized around a general strategy that flows directly from our analysis: *preserve optionality*. The deepest danger of the current moment is not that regulators will choose the wrong policy. It is that, under cover of uncertainty, they will default to inaction that forecloses the ability to choose at all. Preserving optionality means ensuring that regulators retain the capacity to act—in either direction—as the technology's risks and benefits become clearer.

In designing the proposals that follow, we draw on the work of scholars who have developed frameworks for dynamic regulation in other domains characterized by rapid change and deep uncertainty. Environmental law scholar Justin Pidot's helpful taxonomy of "dynamic law" identifies three modes of legal adaptability: durational regulation (rules with built-in expiration dates), adaptive regulation (rules that include processes for ongoing updating), and contingent regulation (rules that automatically activate when predefined triggers are met).[301] Our proposals employ all three. We also draw on Bennear and Wiener's insight that sequential regulatory decisions with monitoring can converge on more optimal outcomes than a single irrevocable choice made under conditions of ignorance,[302] and on the substantial literature applying similar concepts to environmental, nuclear, and dual-use technology governance.[303]

We offer five proposals for present-day regulation directed at addressing existential and catastrophic AI risks. Across all of these, the goal is to ensure that the legal system retains the tools to respond

---

[301] Justin R. Pidot, *Governance and Uncertainty*, 37 CARDOZO L. REV. 112 (2015).

[302] Lori S. Bennear & Jonathan B. Wiener, *Adaptive Regulation: Instrument Choice for Policy Learning over Time* (Draft Working Paper, Revised Feb. 12, 2019), https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/files/Regulation%20-%20adaptive%20reg%20-%20Bennear%20Wiener%20on%20Adaptive%20Reg%20Instrum%20Choice%202019%2002%2012%20clean.pdf.

[303] *See, e.g.*, J.B. Ruhl & Robert L. Fischman, *Adaptive Management in the Courts*, 95 MINN. L. REV. 424 (2010); Daniel Sarewitz, *Governance of Dual-Use Technologies: Theory and Practice*, AM. ACAD. ARTS & SCIS. (2016), https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice; *Periodic Safety Review for Nuclear Power Plants*, IAEA (2013), https://www-pub.iaea.org/MTCD/Publications/PDF/Pub1588_web.pdf.

effectively as the uncertain future trajectory of AI technology becomes clearer.

A. If/Then Regulations

Much of the debate around AI safety centers on when, if ever, AI will develop capabilities sufficient to pose existential threats to humanity. Reasonable people disagree sharply about these timelines.[304] If/then regulations have the enormous benefit of largely sidestepping this disagreement.[305] They impose requirements only if AI achieves a specified capability, taking the general form: if an AI model develops capability X, risk mitigations Y must be in place before it can be further developed or deployed.[306]

Consider an example. If an AI model demonstrates the ability to guide a novice through constructing a weapon of mass destruction, developers should be required to provide assurance that their systems block such attempts, or their systems will be withdrawn from the market.[307] If the skeptics are correct, and AI systems provide little uplift relative to a Google or library search, then these contingencies will not be triggered. Or consider a case where agent swarms are identified as collecting significant resources for unknown reasons without human oversight. In such cases, if-then regulations can trigger, allowing enforcement agencies to audit and, if necessary, order a shutdown of the swarm.

The virtue of if/then structures is that they impose few costs on developers whose systems have not reached the specified capability thresholds or caused the specified harm. As contingent rules, they are likely to be more politically feasible than present-day burdens on AI developers. At the same time, they ensure that safety measures are already in place and immediately enforceable when dangerous capabilities emerge. The alternative, scrambling to develop and enact

---

[304] *Compare, e.g.*, Dean, *supra* note 16 *with* Bengio, et al., *Managing extreme AI risks*, *supra* note 15.

[305] *See* Christopher Painter, *Challenges in Creating "If-Then Commitments" for Extreme Security*, FAR.AI (Feb. 9, 2025), https://www.far.ai/events/sessions/christopher-painter-challenges-in-creating-if-then-commitments-for-extreme-security.

[306] Holden Karnofsky, *If-Then Commitments for AI Risk Reduction*, CARNEGIE ENDOWMENT FOR INT'L PEACE (Sep. 13, 2024), https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction?lang=en; Yoshua Bengio et al., *Managing Extreme AI Risks Amid Rapid Progress*, 384 SCIENCE 843, 844–45 (2024), https://www.science.org/doi/abs/10.1126/science.adn0117.

[307] *Id.*

regulations after a capability has been deployed at scale, is notoriously difficult and likely to be ineffective.[308]

Difficulties will arise in defining the triggering conditions and in determining when they have been activated. But these difficulties are not unique to AI. Environmental regulation routinely ties regulatory requirements to measured concentrations of pollutants;[309] nuclear regulation ties safety requirements to reactor power levels and fuel enrichment thresholds;[310] financial regulation ties enhanced oversight to asset levels and interconnectedness metrics.[311] In each case, the trigger is imperfect, subject to gaming, and requires expert judgment to apply. It is nonetheless far preferable to having no trigger at all. Regulators with sufficient access to AI development processes, aided by internal compliance officers and independent technical evaluators, should be able to identify the emergence of new threats with reasonable effectiveness.[312]

B. Too AI to Fail

The central lesson of the 2008 financial crisis was not just that large institutions can fail. It was that some institutions become so deeply embedded in the broader system that regulators cannot afford to let them fail. When the government concluded that AIG's collapse would bring down the global financial system, it had no choice but to extend a massive bailout, effectively rewarding the recklessness that created the crisis.[313] Today, federal regulators impose enhanced oversight and regulatory requirements for banks and other financial service providers that are too big or too interconnected to let fail, attempting to prevent financial collapse ex ante rather than engaging in costly fixes ex post.[314]

---

[308] *See, e.g.*, GARY E. MARCHANT, THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND THE LAW, IN THE GROWING GAP BETWEEN EMERGING TECHNOLOGIES AND LEGAL-ETHICAL OVERSIGHT: THE PACING PROBLEM 19 (Gary E. Marchant, Braden R. Allenby & Joseph R. Herkert eds., Springer 2011); DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY 11–19 (St. Martin's Press 1980).

[309] *See, e.g.*, Clean Air Act, 42 U.S.C. § 7409.

[310] *See Classification of Uranium*, INT'L ATOMIC ENERGY AGENCY (2022), https://www.iaea.org/newscenter/news/classification-of-uranium (describing uranium enrichment categories, including commonly used thresholds such as 20% U-235).

[311] Dodd-Frank Wall Street Reform and Consumer Protection Act, 12 U.S.C. §§ 5323, 5365.

[312] Karnofsky, *supra* note 306.

[313] *See* MAURICE R. GREENBERG & LAWRENCE A. CUNNINGHAM, THE AIG STORY 250–51 (2013).

[314] MARC LABONTE, CONG. RSCH. SERV., R42150, SYSTEMICALLY IMPORTANT OR "TOO BIG TO FAIL" FINANCIAL INSTITUTIONS (2018).

AI systems embedded in critical infrastructure present some of the same problems for regulators. Such systems may become so systemically important that their failure would threaten essential services or economic stability. This is especially likely if a single AI system is running or interacting with multiple critical infrastructural systems, such that its failure could lead to failures in other, previously unaffected infrastructures.[315] Preventing catastrophic or even existential harms for critical system failures should be a primary goal of regulatory policy.

Moreover, the catastrophic consequences of failing AI systems embedded in critical infrastructure may make it difficult to effectively deter risky behavior via ex-post remedies. AI developers are unlikely to have sufficient assets to compensate for the full social costs of a catastrophe, and the government is likely to bear the enormous costs of disaster at taxpayer expense. As a result, companies will fail to fully internalize the risks that AI system failures pose. Preventative regulation is one way to mitigate the moral hazard before the crisis rather than after.

To address these interlocking risks, systemically important AI systems should be subject to enhanced oversight requirements, including *regular stress testing* to evaluate system behavior under adverse conditions. Such testing should encompass adversarial attacks, out-of-distribution inputs, correlated failures across interconnected systems, and the sudden loss of key components or data feeds. These stress tests should be conducted by independent evaluators, not solely by the systems' developers or operators.[316]

In addition, systemically important companies should be required to establish *mandatory backup systems* that are less dependent on AI and can maintain essential functions if the primary AI system fails. The principle of defense-in-depth, long established in nuclear safety, applies with equal force here: no single system should be a point of failure for critical infrastructure.[317] Safety-oriented design of

---

[315] *See id; supra* Part II.B.3.

[316] For an example from the financial context, *see* 12 U.S.C. § 5365(i); Supervisory and Company-Run Stress Test Requirements for Covered Companies, 77 Fed. Reg. 62,378 (Oct. 12, 2012), https://www.federalregister.gov/documents/2012/10/12/2012-24987/supervisory-and-company-run-stress-test-requirements-for-covered-companies.

[317] *See Defense-in-Depth*, U.S. NUCLEAR REGUL. COMM'N, https://www.nrc.gov/reading-rm/basic-ref/glossary/defense-in-depth.html (last visited ___); INT'L ATOMIC ENERGY AGENCY, IAEA NUCLEAR SAFETY AND SECURITY GLOSSARY, 51.

important infrastructure should address directly the risk of correlated accidents.

Finally, systemically important companies should be subject to *tiered regulatory requirements* calibrated to the criticality and interconnectedness of the AI system in question. More stringent requirements should apply to systems whose failure could affect multiple infrastructure domains or large populations.

The costs of these "too AI to fail" regulatory burdens may be substantial. But in their absence, companies will systematically underprice the downside risks of AI system failures in critical infrastructure. By imposing requirements only on AI systems embedded in critical infrastructure or performing critical functions, and by tiering regulatory requirements based on the essentiality of the systems, the regulatory system is unlikely to overdeter, and firms can easily predict the level of regulatory burden they will incur based on which critical public services they provide.

C. Information Mechanisms

A precondition for effective governance of any kind is information. Regulators cannot act on risks they cannot see. Disclosure requirements related to safety are a relatively simple intervention, and they create an essential foundation for every other proposal in this Part. The marginal cost of federal disclosure requirements would likely be modest. Frontier AI developers already face reporting obligations under California's SB 53 and New York's RAISE Act.[318] A federal disclosure framework that harmonizes and extends these requirements would add relatively little to what major labs must already produce.

The specific content of mandatory disclosures should be developed through notice-and-comment rulemaking, but three broad categories are fundamental. First, mandatory disclosures should include identity and infrastructure disclosures, meaning disclosures of the entities developing frontier models, the locations and computational resources involved in training, and the organizational structures governing safety decisions. Second, they should include safety testing and incident reports, which disclose the results of internal safety evaluations, documented instances of unexpected or dangerous model

---

[318] *See, e.g.*, Gluck, *supra* note 63; Johnson, et al., *supra* note 65.

behavior during development, and any post-deployment incidents meeting defined severity thresholds. These reporting systems should be complemented by robust whistleblower protections for employees who identify safety concerns internally, as California's SB-53 already provides.[319] Third, they should include risk mitigation plans, reporting the measures developers have adopted or intend to adopt to address identified risks. Aviation safety owes much of its remarkable record to mandatory incident reporting to the FAA and the National Transportation Safety Board.[320] An analogous AI incident reporting regime would provide regulators with real-time information about emerging risks, enabling them to act quickly in the event of imminent harms. Further, it would create a shared knowledge base from which the entire field can learn.[321]

D. Deep Learning Regulations

A regulatory framework for AI should be designed, at least in part, to evolve, because the technology it governs will evolve. We propose four adaptive mechanisms, each drawn from established regulatory practice and tailored for the AI context.

Sunset clauses for otherwise static AI regulations could enhance flexibility by requiring AI-specific legislation to expire after a defined period unless affirmatively renewed by Congress. Sunset provisions force periodic reassessment of whether regulatory measures remain appropriate in light of technological change.[322] They address a legitimate concern that regulations designed for today's technology

---

[319] Gluck, *supra* note 63.

[320] *See* FEDERAL AVIATION ADMINISTRATION, AERONAUTICAL INFORMATION MANUAL §7 (Jan. 22, 2026), https://www.faa.gov/air_traffic/publications/atpubs/aim_html/chap7_section_7.html.

[321] Beyond disclosure, the federal government needs technical capacity to make use of the information it receives. This means staffing a dedicated AI safety office with engineers, computer scientists, and domain experts capable of evaluating frontier models and assessing risks. The Nuclear Regulatory Commission employs physicists and engineers who can independently evaluate reactor safety claims; the FDA employs physicians and pharmacologists who can assess drug trial data. *See Office of Nuclear Regulatory Research*, U.S. NUCLEAR REGUL. COMM'N, https://www.nrc.gov/about-nrc/organization/resfuncdesc (last visited ___) (describing NRC technical research and safety-analysis functions supporting regulatory decisions); *Development & Approval Process: Drugs*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/drugs/development-approval-process-drugs (last visited ___) (describing the FDA's technical evaluation role in assessing submissions). An AI safety agency requires comparable technical depth.

[322] Nathan Cortez, *Regulating Disruptive Innovation*, 29 BERKELEY TECH. L.J. 175, 218 (2014).

will become obsolete or counterproductive as capabilities advance.[323] For AI regulation, where the pace of technological change is rapid and uncertainty is high, the case for built-in expiration dates for at least some direct regulation is strong.

Sunrise Clauses operate in the opposite direction: they announce in advance that certain requirements will take effect at a future date, giving industry time to prepare while signaling the direction of regulatory travel.[324] A sunrise clause might provide, for example, that two years after enactment, all AI systems deployed in specified high-risk domains must meet defined safety certification requirements. This gives developers a clear compliance timeline and incentivizes investment in safety research and infrastructure before the requirements become binding. If, over time, we learn that risks were exaggerated, such clauses can be deferred or removed from the books without ever taking effect.

Mandatory periodic review would require designated federal agencies to reassess AI safety regulations at defined intervals (we suggest every three years) in light of new evidence about AI capabilities, failure modes, and the effectiveness of existing measures. The Clean Air Act's requirement that the EPA review National Ambient Air Quality Standards every five years provides a well-established model.[325] Periodic review ensures that regulatory frameworks remain current without requiring the full legislative process for each adjustment. It also creates predictable intervals at which civil society, industry, and technical experts can provide input on whether regulations should be tightened, loosened, or restructured.

Dynamic performance standards would set regulatory requirements in terms of outcomes rather than specific technical means, allowing the required level of performance to adjust as technology advances.[326] Rather than prescribing that AI systems must use a particular safety technique (which may become obsolete),

---

[323] *See id.*; Alasdair Phillps-Robins & Scott Singer, *The State of State AI Law: What's Coming Now that the Federal Moratorium Is Dead*, CARNEGIE ENDOWMENT FOR INT'L PEACE (July 10, 2025), https://carnegieendowment.org/research/2025/07/state-ai-law-whats-coming-now-that-the-federal-moratorium-is-dead.

[324] See AI Act, EUR. COMM'N (Jan. 27, 2026), https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (describing phased applicability dates and delayed obligations for different regulatory modules).

[325] Clean Air Act, 42 U.S.C. § 7409(d)(1).

[326] *See* Cary Coglianese, *Performance-Based Regulation: Concepts and Challenges*, *in* COMPARATIVE LAW AND REGULATION: UNDERSTANDING THE GLOBAL REGULATORY PROCESS (Francesca Bignami & David Zaring, eds., 2016)

regulatory standards could be pegged to broader safety goals. This approach has roots in tort law, where the negligence standard already requires actors to meet the level of care that evolves with industry practice.[327] Dynamic performance standards formalize this logic on the regulatory side, setting explicit benchmarks that adjust as the state of the art advances. For AI, dynamic standards might require, for example, that autonomous systems deployed in transportation maintain accident rates at or below a specified fraction of human-operated baselines, with that fraction decreasing as AI capabilities improve.

Together, these mechanisms create a regulatory framework that can learn and adapt. They address the concern—legitimate in itself but often deployed as a pretext for inaction—that we do not yet know enough about AI and how society will use it to avoid regulatory mistakes.[328] Adaptive mechanisms acknowledge regulators' imperfect knowledge while refusing to treat it as a reason for paralysis.

E. Positive Tax Incentives

The proposals above operate primarily through regulatory requirements. But the market failure we documented in Part II, where AI labs invest tens to hundreds of times more in making their systems powerful than in making them safe, is better addressed by a positive incentive.[329]

Governments can use such incentives to align private investment with public safety goals. This approach has deep roots in American regulatory practice. Tax credits for renewable energy research accelerated the development of solar and wind technology;[330] credits for electric vehicles shifted both consumer purchasing behavior and manufacturer investment;[331] the Orphan Drug Act's tax incentives

---

[327] *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: PHYS. & EMOT. HARM § 13 (2010).

[328] *See, e.g.*, Arvind Narayanan and Sayash Kapoor, *AI existential risk probabilities are too unreliable to inform policy*, AI AS NORMAL TECHNOLOGY, SUBSTACK (July 26, 2024), https://www.normaltech.ai/p/ai-existential-risk-probabilities (arguing that too much about AI is unknown such that the state should not act to regulate it).

[329] *See supra* Part II.C.4.

[330] *E.g.*, Robert Sussman, *Designing the New Green Deal: Where's the Sweet Spot?*, 49 ENVT'L L. REP. NEWS & ANALYSIS 10428, 10445 (2019).

[331] *See* Elliot Wiley, *Pedal into the Future*, 22 SUSTAINABLE DEV. L. & POL'Y 20 (2022).

spurred pharmaceutical research into rare diseases that the market would otherwise have ignored.[332]

Recently, one of us, along with Mirit Eyal, developed a detailed framework applying this principle to AI safety, which we discuss here as an illustration of the approach.[333] This proposal leverages the tax code through three mechanisms: enhanced tax credits for qualifying AI safety research, consumer-side incentives for AI products that meet defined safety certification standards, and recapture provisions that claw back tax benefits from developers whose systems subsequently cause serious harms or fail to meet evolving safety benchmarks.[334] The core insight is that tax incentives can make safety research profitable without making capability research unprofitable, avoiding the zero-sum framing of the race paradigm. In this framework, winning the AI competition means deploying systems that are not merely powerful but demonstrably safe, and the tax code is the mechanism that makes this redefinition stick.[335]

F. Against Regulatory Futility

Critics of AI regulation may argue that unilateral American restraint is futile because other countries will simply develop dangerous AI systems regardless. This argument, a cousin of the "AI arms race" argument, rests on assumptions that do not survive scrutiny. In our previous work with Albert Lin, we have addressed the international context of AI regulation in detail.[336] Above, we have critiqued the concept of an AI arms race between nations.[337] Here we will briefly explain why we believe that American regulation of AI, especially flexible regulation, will promote international cooperation and harmonization rather than local disadvantage.

For a start, the regulatory futility argument assumes that other countries are indifferent to AI safety—an assumption contradicted by the evidence. China has implemented AI regulations, participates in

---

[332] *See, e.g.*, *The Orphan Drug Act: Legal Overview & Policy Considerations*, Library of Congress (Mar. 5, 2024) https://www.congress.gov/crs-product/IF12605; Michael S. Sinha, Ariel D. Stern & Arti K. Rai, *Four Decades of Orphan Drugs and Priorities for the Future*, 391 NEW ENGL. J. MED. 100 (2024).

[333] Mirit Eyal & Yonathan Arbel, *Racing to Safety: Tax Policy for AI Governance*, 79 S.M.U. L. Rev. (forthcoming 2026).

[334] *Id.*

[335] *Id.*

[336] Arbel, et al., *supra* note 18.

[337] *See supra* Part IV.

international AI governance discussions, and has signed international declarations on AI safety that the United States has refused to join.[338] The European Union has enacted the most comprehensive AI regulatory framework in the world.[339] The assumption that Chinese or European policymakers are willing to court civilizational catastrophe for temporary competitive advantage reflects projection, not analysis. Nor is not clear that China perceives itself as engaged in an urgent race with the United States such that it would discard safety for speed, as some policymakers would have us do.[340] China's AI strategy document emphasizes practical applications and industrial upgrading rather than military competition with a named adversary.[341] Chinese leadership appears to expect transformative AI capabilities to arrive later than some American projections suggest, and has focused on using AI "to solve concrete problems" rather than treating it primarily as a strategic weapon.[342] To be sure, China may not wish to broadcast its full intentions, and its posture may shift. But we should at least entertain the possibility that no other runner is sprinting—that Chinese leadership views AI development primarily through an economic lens rather than as a winner-take-all contest.

In addition, successful precedents demonstrate that international coordination on catastrophic risks is achievable even under conditions of intense strategic or economic competition. The Nuclear Non-Proliferation Treaty was negotiated at the height of the Cold War.[343] The Montreal Protocol on ozone depletion achieved near-universal participation despite significant economic costs to affected

---

[338] *See, e.g.*, Laney Zhang, *China: Generative AI Measures Finalized*, L. LIBR. CONG. (July 19, 2023), https://www.loc.gov/item/global-legal-monitor/2023-07-18/china-generative-ai-measures-finalized; Milmo & Courea, *supra* note 52.

[339] *See AI Act*, EUR. COMM'N (updated Jan. 27, 2026), https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[340] *See, e.g.*, Hegseth Memorandum, *supra* note 8, at 4.

[341] *See Global AI Governance Action Plan*, MINISTRY OF FOREIGN AFFAIRS PEOPLE'S REPUBLIC OF CHINA (July 26, 2025), https://www.mfa.gov.cn/eng/xw/zyxw/202507/t20250729_11679232.html.

[342] *Xi Jinping's plan*, *supra* note 36.

[343] *See* Treaty on the Non-Proliferation of Nuclear Weapons, opened for signature July 1, 1968, 729 U.N.T.S. 161 (entered into force Mar. 5, 1970); *Treaty on the Non-Proliferation of Nuclear Weapons (NPT)*, U.N. OFF. DISARMAMENT AFFS., https://disarmament.unoda.org/en/our-work/weapons-mass-destruction/nuclear-weapons/treaty-non-proliferation-nuclear-weapons (last visited ___);

industries.[344] The Biological Weapons Convention was signed by both superpowers despite its implications for military capability.[345] None of these regimes is perfect. But each demonstrates that nations can cooperate to constrain dangerous technologies when the risks are sufficiently clear and the institutional mechanisms are sufficiently credible.

Further, American leadership on AI safety could catalyze rather than hinder international cooperation. Unilateral American safety measures would demonstrate that advanced AI development is compatible with meaningful risk management, potentially encouraging similar measures elsewhere. The United States led in developing nuclear safety standards that became global benchmarks through the IAEA; it could play an analogous role in AI.[346] Conversely, American refusal to implement safety measures provides both motivation and convenient justification for other countries to pursue reckless development trajectories.[347]

The regulatory futility argument ultimately counsels despair. It assumes that, because perfect coordination is unattainable, no coordination is worth attempting—a standard that, if applied consistently, would have precluded every arms control agreement, environmental treaty, and public health convention in history. Even if international coordination proves difficult, domestic safety measures remain valuable for reducing risks from American AI development, which could threaten Americans regardless of what other nations choose to do.

CONCLUSION

This Article has brought the debate over AI existential risk into the mainstream of legal scholarship and demonstrated that it belongs there. We began by documenting a paradox: at the very moment AI

---

[344] *See* Montreal Protocol on Substances that Deplete the Ozone Layer, Sept. 16, 1987, 1522 U.N.T.S. 3) (entered into force Jan. 1, 1989); GABRIELLE DREYFUS, MAX WEI, NIHAR SHAH & KRISTEN TADDONIO, AMBITIOUS REPLENISHMENT OF THE MULTILATERAL FUND FOR THE IMPLEMENTATION OF THE MONTREAL PROTOCOL DELIVERS HIGH IMPACT CLIMATE BENEFITS AT LOW COST 3 (2023).

[345] *See* Convention on the Prohibition of the Development, Production and Stockpiling of *Bacteriological* (Biological) and Toxin Weapons and on Their Destruction, opened for signature Apr. 10, 1972, 1015 U.N.T.S. 163 (entered into force Mar. 26, 1975).

[346] *See The International Atomic Energy Agency*, U.S. DEP'T OF STATE, https://www.state.gov/iaea, (last visited ___).

[347] *See* Jervis, *supra* note 287; Garfinkel & Dafoe, *supra* note 288.

systems are transitioning from chatbots to autonomous agents capable of executing complex objectives in the physical world, the United States has chosen to dismantle even the modest safety regulations that once existed.

To understand the debate around existential risks, we mapped the risk landscape across three categories: human-directed harms, systemic accidents, and loss of control. After engaging seriously with both proponents and skeptics, we showed that, wherever one lands on questions of probability, the risks are sufficiently plausible and their potential consequences sufficiently catastrophic to clear the threshold for regulatory action. We addressed widespread concerns about the effects of regulation on America's competitiveness in the global "AI race" by offering a comprehensive critique of the race metaphor itself. We showed that the metaphor fails as a description of AI competition, which has no stable finish line and is characterized by knowledge diffusion and catch-up dynamics. Further, the concept of a race threatens to corrode safety norms, facilitate proliferation of dangerous weapons, and foreclose cooperation that would make all parties safer.

Based on this analysis, we offered concrete regulatory prescriptions, drawing on the substantial literature on regulation under uncertainty. Our proposals are not a bet on any particular forecast of AI's future. They are an architecture of preparedness. The key directive across all of these is to preserve optionality--to build, *now*, the institutional infrastructure to respond to dangers that may materialize, while remaining flexible and avoiding premature overregulation.

The errors that drive the current deregulatory posture are understandable. AI is developing rapidly, its trajectory is genuinely uncertain, and the economic and strategic stakes are enormous. Policymakers are not wrong to worry about over-regulation. But they are wrong to conclude that uncertainty counsels inaction. The history of catastrophic risk—nuclear, environmental, economic—shows that the worst outcomes do not wait for scientific consensus. The nations and institutions that fare best are those that build the capacity to respond before the crisis arrives. AI's potential harms are plausible, catastrophic, and in many cases permanent. We can recover from imperfect regulation. We cannot recover from extinction.