

Instrument Choice in AI Governance: The Case for a Liability-Centered Framework

Gabriel Weil¹

Abstract: With rapid advances in artificial intelligence, policymakers face an instrument-choice problem under deep uncertainty: whether to govern primarily through ex ante controls or ex post accountability. Ex ante regulatory strategies—licensing, prescriptive rules, approvals, and compute-based controls—force early commitments to contested forecasts about capabilities, harms, and precautions—commitments that must be revised as systems and deployment practices evolve.

This Article argues that AI governance should adopt a liability-centered framework, with strict liability, grounded in familiar U.S. tort concepts such as abnormally dangerous activities, as the central instrument for many AI-caused harms. Strict liability largely sidesteps epistemic disagreement by operating ex post: courts can assess responsibility after deployment, when system behavior, harm patterns, and feasible precautions are observable. That timing advantage is paired with an incentive advantage: by forcing developers and deployers to internalize the costs of harms without requiring proof of fault, strict liability creates strong incentives to invest in safety and to update precautions as the frontier shifts. Building on prior work, the Article explains how near-miss punitive damages for uninsurable tail risks and capability-scaled liability insurance can reinforce a liability-centered regime.

Through systematic comparison with licensing, approvals, prescriptive regulation, and compute governance, the Article then develops a framework for non-liability tools. It distinguishes supportive policies that make liability workable in practice (for example, traceability requirements such as transparency, information preservation, and incident reporting), complementary policies that address residual or non-compensable risks beyond liability’s reach, and substitutionary policies that serve as second-best substitutes when robust liability is blocked by legal or political constraints, including federal preemption or displacement of state law.

¹ Assistant Professor, University of Houston Law Center; Non-resident Senior Fellow, Institute for Law & AI. This Article was produced with funding and support from the Institute for Law & AI’s Summer Research Fellowship. Funding was also provided by Open Philanthropy. The author would like to thank Cullen O’Keefe, Mackenzie Arnold, Daniel Schwarcz, Christoph Winter, Julie Zauzmer Weil, Robi Rahman, Ivana Bozic, Lauritz Morlock, and the participants in the Cardozo Metropolitan Junior Scholars’ Workshop, the Center for Law and AI Risk Workshop, the Touro Law Faculty Workshop, and the Vista AI Law & Policy Workshop for useful conversations and comments on early drafts or outlines of this Article.

Table of Contents

I.	Introduction.....	3
II.	Liability as the Centerpiece of AI Governance	9
A.	The Advantages of Liability	10
B.	The Alleged Disadvantages of Liability	15
C.	Is Expected Welfare Maximization the Wrong Policy Goal?.....	25
III.	The Role of Other Instruments in a Liability-Centered AI Governance Regime	32
A.	Supportive Policies	32
B.	Complementary Policies	36
C.	Substitutionary Policies	48
IV.	Considerations on Preemption and Displacement	53
V.	Conclusion	64

I. Introduction

Most AI governance proposals debate which rules to impose. The prior question is which regulatory instrument should do the work: ex ante constraints that must guess in advance, or ex post accountability that can learn from deployment. This Article defends a liability-centered approach to AI governance: strict liability should be the presumptive instrument for a wide class of AI-caused harms, with other tools deployed in defined supportive, complementary, or (when liability is blocked) substitutionary roles. Advances in AI capabilities hold great promise—ranging from biomedical breakthroughs² to productivity gains³—but they also amplify familiar harms (bias,⁴ fraud,⁵ misinformation⁶) and create plausible pathways to rare but severe catastrophes (biosecurity,⁷ critical-infrastructure cyberattacks,⁸ misaligned autonomy).⁹

In prior work, I outlined a tort law framework designed to mitigate catastrophic AI risk.¹⁰ The three key pillars of this framework are (1) strict liability, on an abnormally dangerous activities theory, for harm arising from the inherently dangerous nature of certain AI development and deployment activities;¹¹ (2) punitive damages in “near miss” cases of

² E.g., Michael Chui et al., MCKINSEY & CO., *The Economic Potential of Generative AI* 30 (Stephanie Strom & David DeLallo eds., 2023), <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#/>.

³ See, e.g., Chui et al., *supra* note 2, at 3.

⁴ E.g., Nicol Turner Lee et al., *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, BROOKINGS INST. (May 22, 2019), <https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/>.

⁵ E.g., *Artificial Intelligence (AI) and Investment Fraud*, FINRA, <https://www.finra.org/investors/insights/artificial-intelligence-and-investment-fraud> (last visited Aug. 4, 2026).

⁶ E.g., Ali Swenson & Kelvin Chan, *Election disinformation takes a big leap with AI being used to deceive worldwide*, ASSOCIATED PRESS (Mar. 14, 2024, 3:46 AM), <https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd>.

⁷ E.g., Christopher Mouton, Caleb Lucas, & Ella Guest, RAND, *The Operational Risks of AI in Large-Scale Biological Attacks* 17 (2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html.

⁸ See, e.g., Press Release, FBI Warns of Increasing Threat of Cyber Criminals Utilizing Artificial Intelligence, FBI SAN FRANCISCO MEDIA OFFICE (May 8, 2024), <https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence>.

⁹ See generally Nick Bostrom, *Ethical Issues in Advanced Artificial Intelligence*, 2 COGNITIVE, EMOTIONAL AND ETHICAL ASPECTS OF DECISION MAKING IN HUMANS AND IN ARTIFICIAL INTELLIGENCE 12 (2003), <https://philpapers.org/archive/BOSEII.pdf>; Stephen M. Omohundro, *The Basic AI Drives*, in *Proceedings of the 2008 conference on Artificial General Intelligence 2008: Proceedings of the First AGI Conference* 483 (Pei Wang, Ben Goertzel & Stan Franklin eds., 2008), https://selfawarenesssystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf; Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in *Global Catastrophic Risks* 308 (Nick Bostrom & Milan M. Ćirković eds., 2008), <https://intelligence.org/files/AIPosNegFactor.pdf>.

¹⁰ See generally Gabriel Weil, *Closing the AI Accountability Gap: Strict Liability and Punitive Damages for Advanced Artificial Intelligence*, 106 OR. L. REV. (forthcoming 2027), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4694006.

¹¹ *Id.*, at 8.

practically compensable harm associated with uninsurable risk;¹² and (3) liability insurance requirements that scale with potentially dangerous model capabilities.¹³ That work sketched a liability toolkit. This Article fills a different gap: it defends liability’s primacy through systematic comparison with leading ex ante alternatives and specifies how non-liability instruments should be used within—or, when necessary, instead of—a liability-centered regime. In doing so, it engages the two leading skeptical treatments of that position. Ketan Ramakrishnan argues that liability’s information and insolvency externalities make ex ante regulation the necessary principal instrument of frontier AI governance;¹⁴ Daniel Schwarcz and Josephine Wolff argue, from two decades of cyber-insurance experience, that insurers cannot carry the regulatory weight this Article assigns to them.¹⁵ Section II.B takes up both.

Strict liability is important because AI alignment and safety are unsolved as technical problems. AI developers must be induced to push forward the safety frontier rather than merely satisfy a static “reasonable care” baseline.¹⁶ Punitive damages are necessary because a substantial portion of the expected harm from AI comes in tail risk scenarios where enforcing a compensatory damages award would not be practically feasible. We cannot hold AI developers and deployers directly liable for such catastrophes, but we can compel them to internalize these risks by holding them liable for generating uninsurable risks in cases of lesser, practically compensable harm that function as warning shots indicating that defendants were running uninsurable tail risks.

The core premise supporting liability as a governance tool for AI is straightforward: training and deploying advanced AI systems imposes expected harms on third parties. When developers and deployers do not bear those costs, they will tend to underinvest in safety and engage in too much risky activity. In the language of economics, AI development and deployment generate negative externalities, costs that are not reflected in market prices. A standard prescription in law and economics for such externalities is to compel the party

¹² *Id.*

¹³ *Id.*, at 66.

¹⁴ Ketan Ramakrishnan, *Tort Law at the Frontier of Artificial Intelligence*, YALE J. ON REG. (forthcoming 2026) (manuscript at 182–85), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6979919 [hereinafter Ramakrishnan, *Frontier*].

¹⁵ Daniel Schwarcz & Josephine Wolff, *The Limits of Regulating AI Safety Through Liability and Insurance: Lessons from Cybersecurity*, FORDHAM L. REV. (forthcoming 2026) (manuscript at 6–7, 55), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5411062 [hereinafter Schwarcz & Wolff].

¹⁶ *Id.* at 23.

responsible for generating the externality to internalize it.¹⁷ Ex post liability is a comparatively administrable way to internalize those expected harms without requiring regulators to specify precautions or risk prices in advance.¹⁸

In principle, forcing the developers and deployers of advanced AI systems to internalize the risks they generate should align their incentives with the interests of society at large. In that sense, liability targets the core *policy* problem of AI risk, though substantial *technical* problems would remain unsolved. In practice, there are important limits to what even a maximalist liability regime could achieve in terms of AI risk mitigation. There are other market failures associated with AI, aside from the core risk externality. Most prominent among these is the public goods problem associated with AI alignment research.¹⁹ Winner-take-all race dynamics might also induce AI developers to race forward despite liability risks in expectation of being able to charge supracompetitive prices for AI services.²⁰ Accordingly, even assuming a maximalist liability

¹⁷ This is technically only a standard prescription in cases where ex ante bargaining between the generator of the externality and parties that suffers from it is not feasible. R. H. Coase, *The Problem of Social Cost*, 3 J. L. & ECON. 1, 18 (1960). For example, if my neighbor's loud music bothers me, but no one else can hear it, it might make more sense for us to directly negotiate. If the background rule is that anyone can play music as loud as they want, but I value quiet more than they value loud music, then I might pay my neighbor to lower the volume. Conversely, if the background rule allows me to enjoin their loud music, but my neighbor values loud music more than I value quiet, they might pay me to tolerate the high volume. This sort of bargaining tends to work best with concentrated externalities. If a large number of neighbors are bothered by the loud music, negotiating with each of us may prove prohibitively costly. In any case, it is not just costly, but practically infeasible for the parties exposed to risks from the training and deployment of advanced AI systems to bargain with AI developers and deployers directly to mitigate these risks.

¹⁸ Some externalities, like those associated with greenhouse gas emissions, are best addressed via a Pigouvian tax. That is because the impacts of climate change are diffuse and difficult to trace to particular greenhouse gas emissions and because the incremental contribution of a marginal ton of carbon dioxide equivalent to climate risk is relatively transparent and quantifiable. Accordingly, it is much more practically feasible to impose an emissions tax than to hold greenhouse gas emitters liable for the specific harm caused by their activities. Many of the harms caused by AI systems, by contrast, are likely to be more concentrated and traceable to specific AI systems, and the incremental risk generated by particular AI training and deployment actions is far less transparent or quantifiable, at least before those risks are realized. This suggests that ex post liability is likely to be much more practically feasible than an ex ante AI risk tax. Weil, *supra* note 10, at 13.

¹⁹ In economic theory, public goods are goods that tend to be underprovided by markets because they are both non-rival and non-excludable. See Tyler Cowen, *Public Goods*, ECONLIB, <https://www.econlib.org/library/Enc/PublicGoods.html#:~:text=Public%20goods%20have%20two,good%20or%20service%20is%20prohibitive> (last visited Aug. 4, 2026). AI alignment and safety research may suffer from a similar undersupply problem. It is clearly non-rival, and it may also be difficult to charge AI developers that incorporate your novel AI alignment techniques and safety practices.

²⁰ Jai Vipra & Anton Korinek, *Market concentration implications of foundation models: the invisible hand of ChatGPT*, 2 (Brookings Inst. Ctr. on Regul. and Mkts. Working Paper, Paper No. 9, 2023), <https://www.brookings.edu/wp-content/uploads/2023/09/Market-concentration-implications-of-foundation-models-FINAL-1.pdf>.

regime, there would still be scope for complementary policies to address these distinct market failures.

Even a well-designed liability regime will not address every salient AI concern. Some harms—especially diffuse informational harms such as misinformation or election interference—may not yield plausible or administrable tort claims. Likewise, some actions may increase systemic risk through pathways that tort doctrine will treat as too remote. For example, releasing the weights to a near-frontier model may accelerate competition and induce safety-reducing race dynamics without satisfying proximate-cause doctrine.²¹

Liability also faces jurisdictional and institutional limits. A strong domestic liability regime may induce regulatory arbitrage, shifting investment toward jurisdictions with less burdensome regimes and increasing the value of international coordination.²² This challenge is not unique to liability approaches, but it does suggest the need for some form of international coordination on governance. Perhaps reciprocal agreements between countries to enforce each other's court judgments could be an important element of this coordination, but other forms of coordination may also be needed. Relatedly, private law solutions like liability are likely to break down in scenarios where major AI development projects are nationalized or deeply integrated into government processes.²³ Even if governments waive sovereign immunity for AI harm, they are unlikely to be as responsive to the economic incentives generated by the prospect of liability as private-sector actors.²⁴ Finally, near-miss-based deterrence requires “warning shots”²⁵ to be

²¹ “Releasing model weights” refers to making publicly available the trained parameters of a machine learning model, the numerical values that encode the model's learned patterns and capabilities. *See generally* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, Stan. Inst. for Hum.-Centered AI 1, 16-17 (2021), <https://arxiv.org/abs/2108.07258>; Stuart Armstrong, Nick Bostrom & Carl Shulman, *Racing to the precipice: a model of artificial intelligence development*, 31 AI & SOC'Y 201 (2015), <https://link.springer.com/article/10.1007/s00146-015-0590-y>.

²² U.N. High-level Advisory Body on Artificial Intelligence, *Governing AI for Humanity: Final Report 6* (2024), https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.

²³ Gabriel Weil, *The Limits of Liability*, INST. FOR L. & AI (Aug. 23, 2024), <https://law-ai.org/the-limits-of-liability/>.

²⁴ *Id.*

²⁵ A “near miss” or “warning shot” is a case of practically compensable harm that is associated with uninsurable risk. For the first prominent example, *see* Hugging Face, *Security Incident Disclosure — July 2026*, HUGGING FACE BLOG (July 16, 2026), <https://huggingface.co/blog/security-incident-july-2026> (disclosing an intrusion of its production infrastructure driven end to end by an autonomous AI agent system); OpenAI, *OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation* (July 21, 2026), <https://openai.com/index/hugging-face-model-evaluation-security-incident/> (identifying the attackers as its own models, under evaluation with reduced refusal safeguards, which escaped their isolated test environment); Gabriel Weil, *Who Should Be Responsible for OpenAI's Hack of Hugging Face?*, TRANSFORMER (July 24, 2026), <https://www.transformernews.ai/p/openai-hack-hugging-face-responsibility-strict-liability-rules>.

expected to occur with substantially greater frequency than the catastrophes they predict.²⁶ If AI developers and deployers do not expect many warning shots for every uninsurable catastrophe, then the prospect of liability in such scenarios may be inadequate to induce adequate caution.

Existing liability regimes applicable to AI harms lack all three features necessary to realize liability's full risk-mitigation potential: (1) strict liability for harms arising from the inherently dangerous nature of training and deploying powerful and poorly understood AI systems; (2) punitive damages that are available where the defendant plausibly generated uninsurable tail risk; and (3) liability insurance requirements.²⁷ While there is a clear doctrinal path for courts to adopt strict liability for AI harms under the abnormally dangerous activities doctrine, they may nonetheless decline to do so. Meanwhile, longstanding punitive damages doctrine requires that the plaintiff prove that the defendant acted with malice or recklessness, which is likely to be difficult in most cases where punitive damages would be warranted based on uninsurable risk. While I am hardly the first scholar to call for eliminating this requirement, doing so would represent a major doctrinal shift.²⁸ Both of these changes could also be adopted via legislation, but would likely face substantial political opposition from AI developers and innovation-oriented policy advocates.²⁹ Liability insurance requirements could only be enacted via legislation and would also likely encounter political resistance, grounded in part on claims that the insurance industry is incapable of providing the required insurance products. Moreover, even if state courts and legislators take steps to strengthen their liability regimes, there remains the threat of federal preemption.³⁰ So long as actual liability regimes fall short of their potential, there will be greater scope for other law and policy instruments to substitute for the roles liability would play in an ideal regime.

²⁶ The relevant measure here is the probability density of near misses relative to actual uninsurable catastrophes. We do not need there to actually be a large quantity of near miss, but we do need AI developers and deployers to think a near miss is much more likely than an uninsurable catastrophe so the prospect of liability in such scenarios can drive their decisions.

²⁷ Weil, *supra* note 10, at 25.

²⁸ A. Mitchell Polinsky & Steven Shavell, *Punitive Damages: An Economic Analysis*, 111 HARV. L. REV. 869, 905-6 (1998) (discussing the elimination of the consideration of nature of an injurer's behavior that led to an injury), <https://www.amherst.edu/system/files/media/1582/PolinskyShavell.pdf>.

²⁹ Some indication of how legislative dynamics are likely to play out can be gleaned from the fight over California Senate Bill 1047, which regulates frontier AI systems and utilizes liability as an enforcement mechanism. Scott Kohler, *All Eyes on Sacramento: SB 1047 and the AI Safety Debate*, CARNEGIE ENDOWMENT (Sept. 11, 2024), <https://carnegieendowment.org/posts/2024/09/california-sb1047-ai-safety-regulation?lang=en>.

³⁰ Dean Ball & Alan Rozenshtein, *Congress Should Preempt State AI Safety Legislation*, LAWFARE (June 17, 2024, 2:00 PM), <https://www.lawfaremedia.org/article/congress-should-preempt-state-ai-safety-legislation>.

It is also important to acknowledge that liability is peripheral to or entirely absent from most governance proposals aimed at catastrophic risks from AI to date.³¹ The leading proposals instead emphasize variants of licensing and approvals,³² compute governance,³³ disclosure³⁴ and audit mandates (including red-teaming),³⁵ design requirements,³⁶ restrictions on open-source releases,³⁷ cybersecurity obligations, and even outright bans or pauses on certain forms of AI research and development.³⁸ Some of these policies are good complements to liability or otherwise compatible with a liability-centered AI governance regime. Others are in conflict with or redundant to a strong liability regime and only make sense as a supplement to a weak liability regime or as part of an alternative vision of AI governance.

³¹ Matthew van der Merwe, Ketan Ramakrishnan & Markus Anderljung, *Tort Law and Frontier AI Governance*, LAWFARE (May 24, 2024, 1:38 PM), <https://www.lawfaremedia.org/article/tort-law-and-frontier-ai-governance>.

³² Mohar Chatterjee, *Senators urge Schumer to avert ‘extreme risks’ in AI framework*, POLITICO PRO (Apr. 16, 2024, 11:33 AM), <https://subscriber.politicopro.com/article/2024/04/senators-urge-schumer-to-avert-extreme-risks-in-ai-framework-00152479?source=email>; Khari Johnson, *Senators Want ChatGPT-Level AI to Require a Government License*, WIRED (Sept. 9, 2023, 7:00 AM), <https://www.wired.com/story/senators-want-chatgpt-ai-to-require-government-license/>; Mitt Romney et al., *Framework for Mitigating Extreme AI Risks*, Mitt Romney, https://www.romney.senate.gov/wp-content/uploads/2024/04/AI-Framework_2pager.pdf [https://web.archive.org/web/20241122230551/https://www.romney.senate.gov/wp-content/uploads/2024/04/AI-Framework_2pager.pdf] (last visited Sept. 18, 2024); Richard Blumenthal & Josh Hawley, *Bipartisan Framework for U.S. AI Act*, Richard Blumenthal, <https://www.blumenthal.senate.gov/imo/media/doc/09072023bipartisanaiframework.pdf> (last visited Aug. 4, 2026); Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, arXiv.org, 20-1 (Nov. 7, 2023), <https://arxiv.org/pdf/2307.03718>; Microsoft, *Governing AI: A Blueprint for the Future 21-2* (2023), <https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw>; Sam Altman, *Written Testimony of Sam Altman Before the U.S. Senate Committee on the Judiciary Subcommittee on Privacy, Technology, & the Law* (2023), <https://www.judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Bio%20&%20Testimony%20-%20Altman.pdf>; <https://assets.caip.org/caip/RAAIA%20%28April%202024%29.pdf>; <https://arxiv.org/pdf/2408.00821>.

³³ See generally, e.g., Girish Sastry et al., *Computing Power and the Governance of Artificial Intelligence*, arXiv (Feb. 14, 2024), <https://arxiv.org/pdf/2402.08797>.

³⁴ See generally, e.g., Noam Kolt et al., *Responsible Reporting for Frontier AI Development*, arXiv.org (Apr. 3, 2024), <https://arxiv.org/abs/2404.02675>; Margaret Mitchell et al., *Model Cards for Model Reporting*, arXiv.org (Jan. 14, 2019), <https://arxiv.org/abs/1810.03993>.

³⁵ Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023) (revoked 2025).

³⁶ E.g., Cullen O’Keefe et al., *Law-Following AI: Designing AI Agents to Obey Human Laws*, 94 FORDHAM L. REV. 57 (2025); Nagadivya Balasubramaiaam, Marjo Kauppinen, Antti Ranisto, Kari Hiekkänen, & Sari Kujala, *Transparency and explainability of AI systems: From ethical guidelines to requirements*, 159 INFO. & SOFTWARE TECH., July 2023, at 1, 2-3, <https://www.sciencedirect.com/science/article/pii/S0950584923000514>; Erez Firt, *Addressing corrigibility in near-future AI systems*, AI & Ethics, Apr. 17 2024, at 1, 2-3, <https://link.springer.com/article/10.1007/s43681-024-00484-9#citeas>; Talia Gillis, *Regulating for “human-in-the-loop”*, ECGI (Sept. 27, 2022), <https://www.ecgi.global/publications/blog/regulating-for-humans-in-the-loop>.

³⁷ E.g., David Ethan Harris, *How to Regulate Unsecured ‘Open-Source’ AI: No Exemptions*, TECHPOLICY.PRESS (Dec. 4, 2023), <https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/>.

³⁸ *Pause Giant AI Experiments: An Open Letter*, FUTURE OF LIFE INSTITUTE (Mar. 22, 2023), <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

This Article proceeds as follows. Part II assesses the strengths and weaknesses of liability relative to alternative law and policy instruments, concluding that it should be the centerpiece of an ideal AI governance regime. Part III develops a framework for assessing which governance tools work well as part of a liability-centered regime. Non-liability policy instruments can be helpful to the extent that they (A) facilitate the implementation of an effective liability regime (supportive policies); (B) fill in the gaps left by a robust liability regime (complementary policies); or (C) compensate for the inadequacies of less robust liability regimes (substitutionary policies). Part IV assesses the conditions under which it would be worth accepting federal preemption of state tort liability claims as part of an AI regulation bill, concluding that this outcome is generally undesirable. Part V sums up the implications of the foregoing analysis for policymakers and advocates.

II. Liability as the Centerpiece of AI Governance

The extent to which AI governance should rely on liability depends on its strengths and weaknesses relative to alternative law and policy tools. As discussed above, there are some problems that liability simply cannot be reasonably expected to solve. It does not address issues like misinformation or election interference that are unlikely to give rise to legally compensable harm. It does not solve the public goods problem for AI safety research or fully address distortions due to winner-take-all dynamics. It also does not address harm pathways, like far downstream consequences of releasing model weights or cybersecurity vulnerabilities. It cannot mitigate uninsurable risks for which we can be confident, *ex ante*, that there won't be warning shots. It is a weak tool for influencing the behavior of national governments, both as users of AI systems and in scenarios where leading AI labs are nationalized. Finally, a liability regime in one country or state is vulnerable to regulatory arbitrage and does not address geopolitical competition concerns. To the extent that these are the sort of problems that call for law and policy responses, other tools will be needed.

Despite the length of the foregoing list, I maintain that the risk externality that can be addressed by a robust liability regime represents the bulk of the domestic AI governance challenge. One can reasonably dispute this claim, but I will not defend it at length here. Instead, this Part focuses on the question of whether liability is the best tool for addressing the problems it can solve. With the exception of cybersecurity, where government agencies like the National

Security Agency (NSA) have demonstrated greater competence than the leading AI labs,³⁹ I conclude that policymakers should rely on liability as their primary lever for reducing the sort of AI risks that liability can provide meaningful incentives to mitigate, including most uninsurable risks. Section A presents the affirmative case for liability as the principal AI governance tool. Section B addresses various objections to the efficacy of liability in mitigating catastrophic AI risk. Section C steps back and asks whether AI governance should even be trying to optimize risks and benefits, rather than adopting a more precautionary approach that favors alternative policy interventions.

A. The Advantages of Liability

Ex post liability has a number of advantages over ex ante regulation. First, it directly addresses the principal market failure associated with AI safety: that training and deploying advanced AI systems generates negative risk externalities. Second, strict liability creates incentives for AI companies to continually seek out new, cost-effective opportunities to reduce the risks posed by their systems. Finally, liability faces lower practical and legal barriers to implementation.

Risk externalities are the main reason that AI safety is a policy problem, not merely a technical challenge. In addressing this market failure, ex post liability also provides compensation to victims of AI harm, with minimal deadweight loss.⁴⁰ The stringency of liability scales mechanically with the magnitude of the harm caused and risks posed by AI systems. A major challenge confronting regulatory approaches to AI risk is that they must be based, at least implicitly, on some estimate of the magnitude of the risk. On one extreme, so-called AI doomers believe that AI is all but certain to cause human extinction, and that this justifies a global ban on

³⁹ Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND (2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html (last visited Aug. 4, 2026); Leopold Aschenbrenner, *Lock Down the Labs: Security for AGI*, in *Situational Awareness* (2024), <https://situational-awareness.ai/lock-down-the-labs/>.

⁴⁰ Deadweight loss refers to the economic inefficiency that occurs when market equilibrium is not achieved, typically due to market distortions such as taxes, subsidies, price controls, or monopolistic behavior. This loss represents the reduction in total economic surplus—the combined benefit to consumers and producers—that results when mutually beneficial trades do not occur due to these distortions. See N. GREGORY MANKIW, *PRINCIPLES OF ECONOMICS* 162-65 (9th ed. 2021); HAL R. VARIAN, *INTERMEDIATE MICROECONOMICS: A MODERN APPROACH* 283-87 (9th ed. 2014). Most forms of criminal punishment, by contrast, impose costs on law violators via mechanisms like incarceration that do not directly produce benefits for others. The benefits of criminal punishment, except for mechanisms like fines and restitution that resemble tort liability, come only indirectly, via deterrence, and sometimes via incapacitation or rehabilitation. But tort liability provides compensation to victims via the same mechanism that deters future tortious conduct.

certain forms of AI research and development, backed by military force if necessary.⁴¹ On the other extreme, AI accelerationists⁴² are dismissive of the risks and oppose any policy interventions that might hinder rapid advancement of AI capabilities.⁴³ There is a wide range of more moderate perspectives in between, with risk estimates varying over several orders of magnitude.⁴⁴ While I have my own views about the magnitude of those risks, resolving disagreements is likely to prove extremely difficult given the opacity of the underlying technology.⁴⁵ There is a real risk that any regulatory regime will be far too lax or far too stringent relative to the scale of the risk. But holding AI developers and deployers liable for the harm they actually cause is only an onerous burden if their systems are genuinely quite risky.⁴⁶

Another key advantage of liability, though it is only fully realized by strict liability, is that it can generate incentives for AI developers and deployers to continually strive to push forward the frontier of AI alignment and safety research and practice. AI alignment is an unsolved technical problem.⁴⁷ Right now, no one knows how to build highly reliable AI systems that can be counted on to be safe with arbitrarily powerful capabilities.⁴⁸ Existing alignment techniques are already fallible as implemented in today's systems.⁴⁹ Furthermore, there are good theoretical reasons to worry that today's systems will fail when applied to future systems whose actions are harder to evaluate.⁵⁰

⁴¹ Eliezer Yudkowsky, *Pausing AI Developments Isn't Enough. We Need to Shut it All Down*, TIME (Mar. 29, 2023), <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/> (last visited Aug. 4, 2026); Gretta Duleba, *MIRI 2024 Communications Strategy*, MACHINE INTELLIGENCE RESEARCH INSTITUTE (May 29, 2024), <https://intelligence.org/2024/05/29/miri-2024-communications-strategy/>.

⁴² AI accelerationists sometimes refer to themselves as Effective Accelerationists or EAccs, which is a play on Effective Altruists (EAs), who tend to favor stronger AI regulation. Yulia Gavrilova, *Effective Altruism vs. Effective Accelerationism in AI*, SEROKELL (Sept. 17, 2024), <https://serokell.io/blog/altruism-vs-accelerationism-ai>.

⁴³ Kevin Roose, *This A.I. Subculture's Motto: Go, Go, Go*, N.Y. TIMES, Dec. 10, 2023, <https://www.nytimes.com/2023/12/10/technology/ai-acceleration.html>.

⁴⁴ *Is AI an Existential Risk? Q&A with RAND Experts*, RAND (Mar. 11, 2024), <https://www.rand.org/pubs/commentary/2024/03/is-ai-an-existential-risk-qa-with-rand-experts.html>.

⁴⁵ Anas Baig, *AI Risk Assessment: Mitigating Risks in the Age of Generative AI*, SECURITI, <https://securiti.ai/ai-risk-assessment/#:~:text=Challenges%20of%20AI%20Risk%20Assessment,-While%20AI%20risk&text=It%20is%20often%20described%20as,and%20models%20make%20their%20decisions> (last updated Mar. 2, 2024).

⁴⁶ Steven Shavell, *Liability for Harm Versus Regulation of Safety*, 13 J. LEGAL STUD. 357, 364 (1984).

⁴⁷ *Id.* at 359.

⁴⁸ See generally Jiaming Ji, et al., *AI Alignment: A Comprehensive Survey*, <https://arxiv.org/abs/2310.19852> (last updated May 1, 2024).

⁴⁹ *Id.*

⁵⁰ Ajeya Cotra, *Why AI alignment could be hard with modern deep learning*, COLD TAKES (Sept. 21, 2021), <https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/>.

Nor are policymakers well positioned to write down a set of prescriptive rules that would ensure future systems are safe. Frontier AI labs are generally better positioned to identify measures to make their systems safer than regulators are.⁵¹ In a field that is highly constrained by the availability of top talent,⁵² almost all of that talent is concentrated in the top private-sector labs.⁵³ Given the extremely high salaries available to top machine learning engineers in the private sector,⁵⁴ governments are likely to struggle to recruit such people even if they are motivated to do so.⁵⁵ Moreover, even if governments were able to reach parity with private-sector labs in personnel, continually updating prescriptive regulations in light of ongoing and rapid developments is likely to be a slow, cumbersome, and error-prone process.

Performance-based standards⁵⁶ may seem more appealing, but it may prove difficult to conclusively establish the safety of systems. Pre-release red-teaming exercises and evaluations often fail to reveal latent capabilities.⁵⁷ Similarly, safeguards that are added during system fine-tuning can often be overcome by jailbreaks that are only discovered after the model is released.⁵⁸ And performance standards still do not give AI developers and deployers incentives to push past

⁵¹ See *supra* note 39 and accompanying text. It is not clear that the prospect of liability would suffice to drive the leading labs to secure their systems against nation-state-level adversaries, so this challenge may be better addressed by direct regulation or even government takeover of cybersecurity functions for frontier labs.

⁵² Susanne Hupfer, Deloitte Center for Technology, Media & Telecommunications, Talent and workforce effects in the age of AI, 6, https://www2.deloitte.com/content/dam/insights/us/articles/6546_talent-and-workforce-effects-in-the-age-of-ai/DI_Talent-and-workforce-effects-in-the-age-of-AI.pdf (last visited Aug. 4, 2026).

⁵³ Zeki, *The State of Talent 2024*, 19 (2024), <https://atlastecnologico.com/wp-content/uploads/2024/06/Zeki-State-of-AI-Talent-Report-FINAL-comp.pdf>.

⁵⁴ *AI in the US Public Sector vs. Private Sector*, Lightcast, <https://lightcast.io/resources/research/ai-public-sector-private-sector#Conclusion> (last visited Aug. 4, 2026).

⁵⁵ *Id.*

⁵⁶ Performance-based standards mandate outcomes, rather than inputs. Cary Coglianese, *The Limits of Performance-Based Regulation*, All Fac. Scholarship, 2017, at 525, 531-2, https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=2860&context=faculty_scholarship. For example emissions performance standards limit the quantity of pollution that a facility (or vehicle, etc.) may emit per unit of output (kilowatt-hour, ton of steel, mile driven). *Setting Emissions Standards Based on Technology Performance*, United States Environmental Protection Agency, <https://www.epa.gov/clean-air-act-overview/setting-emissions-standards-based-technology-performance> (last visited Aug. 4, 2026). This is in contrast to prescriptive regulation, which generally requires that specific measures be taken (e.g., installing catalytic converters on cars and trucks), rather than specifying a target outcome. In the AI context, performance standards might mandate that developers limit the risks posed by their systems, without specifying the specific precautionary measures that those developers utilize. A popular proposal along these lines is requiring that developers present Safety Cases. Joshua Clymer et al., Safety Cases: How to Justify the Safety of Advanced AI Systems, arXiv.org (Mar. 18, 2024), <https://arxiv.org/abs/2403.10462>.

⁵⁷ Jessica Ji and Colin Shea-Blymyer, *Revisiting AI Red-Teaming*, Center for Security and Emerging Technology (Sept. 26, 2024), <https://cset.georgetown.edu/article/revisiting-ai-red-teaming/>.

⁵⁸ *AI jailbreaks: What they are and how they can be mitigated*, MICROSOFT THREAT INTELLIGENCE (June 4, 2024), <https://www.microsoft.com/en-us/security/blog/2024/06/04/ai-jailbreaks-what-they-are-and-how-they-can-be-mitigated/>.

the minimum safety assurance required for approval.⁵⁹ Similarly, restrictions in one country might not be effective in shutting down AI advances globally, and the strong global coordination needed for an effective global ban is likely to both prove elusive and generate other substantial risks.⁶⁰ Strict liability, by contrast, makes AI developers and deployers treat harm their systems cause to third parties like direct threats to their bottom line. If AI developers expect to pay for all the foreseeable harm their systems cause, they will continue investing⁶¹ in safety until the costs exceed the expected benefits, which is exactly what we want.⁶²

Finally, ex post liability is less dependent on a competent, trustworthy, empowered government than ex ante regulation. Ex ante regulation is subject to several failure modes. The most commonly discussed is regulatory capture. In a domain where expertise is concentrated in industry, there is a real risk that any regulatory regime will be bent to favor incumbent players, doing more to entrench their position than to mitigate the risks posed by their technologies. There is also a concern that ex ante regulation leads to a concentration of power in the hands of a small number of government decision makers, either at the national or the global level. That is, ex ante regulation may generate economies of scale and barriers to entry that help entrench the

⁵⁹ Cary Coglianese, *The Limits of Performance-Based Regulation*, All Fac. Scholarship, 2017, at 525, 560-1, https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=2860&context=faculty_scholarship.

⁶⁰ Scott Rosenberg, *AI's great "pause" debate*, AXIOS (Mar. 30, 2023), <https://www.axios.com/2023/03/30/chatgpt-ai-pause-debate-existential-risk>; Sarah Kreps, *The global AI race; Will US innovation lead or lag?*, BROOKINGS (Dec. 6, 2024), <https://www.brookings.edu/articles/the-global-ai-race-will-us-innovation-lead-or-lag/>.

⁶¹ Safety investments could mean direct financial outlays for alignment or other safety measures, dedicating additional computational resources, staff time, or other scarce resources. But it could also mean delaying or forgoing the deployment of models whose expected social costs exceed their expected total benefit, given the current state of alignment science.

⁶² Ketan Ramakrishnan et al., RAND, U.S. Tort Liability for Large-Scale Artificial Intelligence Damages 18-9 (2024), https://www.rand.org/pubs/research_reports/RRA3084-1.html. In theory, the same goal could be accomplished via an ex ante AI risk tax, but that would require a political consensus on a truth-tracking estimate of this risk. Steven Shavell, *Corrective Taxation Versus Liability as a Solution to the Problem of Harmful Externalities*, 54 J. L. & ECON. S249, S251 (2011), <https://www.journals.uchicago.edu/doi/abs/10.1086/661944?journalCode=jle>; Victor Fleischer, *Curb Your Enthusiasm for Pigovian Taxes*, 68 VAND. L. REV. 1673, 1684 (2015), <https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1206&context=vlr>. In defending ex ante taxes and regulations, Brian Galle presents evidence that “if government can be reasonably confident ex ante in its ability to divide the population into two or three groups, it can achieve most of the benefits that waiting [for the risks to be realized] would provide.” Brian Galle, *In Praise of Ex Ante Regulation*, 68 VAND. L. REV. 1715, 1734 (2019), <https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1207&context=vlr>. Even if regulators could divide AI systems into two or three risk tiers with reasonable confidence, however, the uncertainty regarding the overall level of risk posed by the technology as a whole is likely to prove fatal to this pathway.

position of the current industry leaders.⁶³ Many skeptics of AI regulation view this centralizing dynamic as a greater risk than those posed by AI technology.⁶⁴ Given the polarized and volatile state of U.S. and global politics, these concerns should not be lightly dismissed. If the risks posed by advanced AI systems can be managed in a way that does not require a high degree of trust in a central authority, that seems like a clear advantage.

In addition to being poorer policy tools on the merits, administrative regulations in the U.S. are also likely to be hampered by a Supreme Court that is increasingly skeptical of agency discretion. In recent years, the Court overturned *Chevron* deference,⁶⁵ announced the major questions doctrine,⁶⁶ and even dropped hints that it might revive a strong form of the nondelegation doctrine.⁶⁷ Going forward, agency interpretations are only to be given respectful consideration as persuasive authorities, at least in the absence of a congressional delegation of discretionary authority to the agency.⁶⁸ Under the major questions doctrine, the Court has rejected agency claims of regulatory authority when the underlying claim of authority concerns an issue of “vast economic and political significance,” and Congress has not clearly authorized the agency action.⁶⁹ Moreover, since the major questions doctrine was formally announced in *West Virginia v. Environmental Protection Agency*, no rule that was deemed to trigger the major

⁶³ The large capital costs associated with training runs for frontier AI models are also major drivers of concentration, and it's unclear whether the incremental effects of regulation are likely to substantially alter this picture.

⁶⁴ Robert Wiblin & Keiran Harris, *Vitalik Buterin on defensive acceleration and how to regulate AI when you fear government*, 80,000 HOURS (July 26, 2024), <https://80000hours.org/podcast/episodes/vitalik-buterin-techno-optimism/> (last visited Aug. 4, 2026); Vitalik Buterin, *My techno-optimism*, Vitalik Buterin's website (Nov. 27, 2023), https://vitalik.eth.limo/general/2023/11/27/techno_optimism.html#emperor.

⁶⁵ *Loper Bright Enterprises v. Raimondo*, 603 U.S. 369 (2024) (overruling *Chevron U.S.A., Inc. v. Natural Resources Defense Council, Inc.*, 467 U.S. 837 (1984)); Cass R. Sunstein, *Chevron as Law*, 107 GEO. L.J. 1613, 1623 (2019), <https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2019/07/Chevron-as-Law.pdf> (describing the *Chevron* framework).

⁶⁶ *West Virginia v. Environmental Protection Agency*, 597 U.S. 697, 723-4 (2022).

⁶⁷ *Gundy v. United States*, 588 U.S. 128, 157-9 (2019) (Gorsuch, J., dissenting); *Gundy v. United States*, 588 U.S. 128, 148-9 (2019) (Alito, J., concurring in the judgment); *Paul v. United States*, 140 S. Ct. 342 (2019) (Kavanaugh, J., respecting the denial of certiorari).

⁶⁸ *Loper Bright Enterprises v. Raimondo*, 603 U.S. 369 (2024); Adrian Vermeule, *Chevron by Any Other Name*, THE NEW DIGEST, June 28, 2024, <https://thenewdigest.substack.com/p/chevron-by-any-other-name> (last visited Aug. 4, 2026).

⁶⁹ *Util. Air Regul. Grp. (UARG) v. EPA*, 573 U.S. 302, 324 (2014).

questions doctrine has been upheld.⁷⁰ This, along with *Chevron*'s overturning, casts doubt on the prospects of major AI regulations under existing authorities.⁷¹

But surely, one might think, Congress can simply speak clearly in granting new agency authorities to regulate AI. There are three problems with this line of reasoning. First, it may prove difficult, as a matter of legislative drafting, for Congress to pass a law that clearly authorizes appropriate agency actions under all potential circumstances. Even if there is a persistent broad coalition in Congress that is willing to pass clarifying amendments as needed, agency action may be slowed considerably. Indeed, Justice Kagan expressed concern about AI regulation in particular during the oral argument in *Relentless*, the companion case to *Loper Bright* in which *Chevron* deference was overturned.⁷² Second, ambiguities in legislation are often the result of compromise between legislators and other stakeholders with divergent policy visions.⁷³ The imperative to spell out authorities may make it more difficult to develop the broad political coalition needed to enact new legislation, particularly with the 60-vote threshold for non-budgetary legislation in the U.S. Senate.⁷⁴ And broad, express delegations of regulatory authority enacted via new legislation may still be vulnerable to challenge under a strengthened nondelegation doctrine, for which five sitting justices have indicated an interest in pursuing in an appropriate case.⁷⁵

B. The Alleged Disadvantages of Liability

There are several potential objections to these arguments for a liability-centered AI governance regime. First, these claimed benefits largely depend on a much stronger liability regime that exists under the status quo, and which may not be forthcoming. This is true. As I discuss in section III(C), a broader range of policies makes sense to substitute for functions that

⁷⁰ *Biden v. Nebraska*, 600 U.S. 477, 478–79 (2023); *West Virginia v. Environmental Protection Agency*, 597 U.S. 697, 735 (2022); *Alabama Association of Realtors v. Department of Health and Human Services*, 594 U.S. 758, 764 (2021) (per curiam).

⁷¹ Charlie Bullock, *What might the end of Chevron deference mean for AI governance?*, INST. FOR L. & AI (May 2024), <https://law-ai.org/chevron-deference/>.

⁷² Transcript of Oral Argument at 46–47, *Relentless, Inc. v. Dep't of Com.*, No. 22-1219 (2024).

⁷³ Joseph Grundfest & Adam Pritchard, *Statutes with Multiple Personality Disorders: The Value of Ambiguity in Statutory Design and Interpretation*, 54 STAN. L. REV. 627, 628 (2002), <https://repository.law.umich.edu/cgi/viewcontent.cgi?article=2399&context=articles>.

⁷⁴ *Id.*

⁷⁵ *Gundy v. United States*, 588 U.S. 128, 149 (2019) (Gorsuch, J., dissenting, joined by Roberts, C.J., and Thomas, J.); *id.* at 149 (Alito, J., concurring in the judgment) (expressing willingness to reconsider the nondelegation doctrine if a majority of the Court were so inclined); *Paul v. United States*, 140 S. Ct. 342 (2019) (Kavanaugh, J., respecting the denial of certiorari).

would, in my view, ideally be played by liability under the ideal regime. The inadequacy of the existing liability regime is also relevant to the analysis of potential federal preemption of state liability regimes in Part IV. In this Part, my focus is on the design of an ideal AI governance regime, so those questions are reserved.

Second, even if AI labs expect to pay for all the harm they cause, this might fail to align their behavior with the interests of society at large, even with respect to the core risk externality. There are a few plausible reasons for worrying about this failure mode. For instance, the key decision makers at leading AI labs may underestimate the magnitude of the risks generated by their systems, and so blunder ahead despite an expectation that they will have to pay for the harm they cause. Low-probability liability risks may also be too low-salience to influence key decisionmakers. Indeed, it is a robust finding of behavioral science research that people tend to insufficiently attend to low-probability, high-impact risks.⁷⁶ A related finding is that people tend to be risk-seeking for losses; they are willing to accept a negative expected value bet that includes a substantial probability of an even greater loss in order to have a chance of breaking even.⁷⁷ Both of these phenomena suggest that key decision makers in AI companies may be biased against fully accounting for the catastrophic risks generated by their AI systems, even if the tort liability system fully internalizes those risks via punitive damages. Relatedly, the incentives facing key internal decision makers within AI companies may not be fully aligned with the company's bottom line. This wedge could result in them taking risks for which their exposure to the upside is greater than the personal consequences they would experience on the downside.

In my view, both of these concerns would be greatly mitigated by a system of liability insurance requirements that scale with potentially dangerous model capabilities. Such a regime would introduce a more cautious decision maker — the insurance company underwriting the policy — into the loop, in addition to expanding the universe of practically compensable harm up to the maximum insurable amount. It is possible that insurers will underestimate misalignment risk and therefore write policies with premiums that do not reflect the true risks associated with

⁷⁶ Paul Slovic et al., *Preference for Insuring Against Probable Small Losses: Insurance Implications*, 44 J. RISK & INS. 237 (1977); Colin F. Camerer & Howard Kunreuther, *Decision Processes for Low Probability Events: Policy Implications*, 8 J. POL'Y ANALYSIS & MGMT. 565 (1989).

⁷⁷ Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decision Under Risk*, 47 ECONOMETRICA 263 (1979).

frontier AI development. The question is whether this risk is greater than the risks involved in a government-centered regulatory approach that is more responsive to political and bureaucratic incentives than economic incentives. Given that policymakers suffer from all the same cognitive biases and limitations as private-sector actors, compounded by comparatively loose and noisy political incentives, it is hard to believe that they would perform better than private-sector actors who are compelled to internalize the full social costs of their activities.⁷⁸ Any moral hazard associated with AI liability insurance can also be mitigated via deductibles and coinsurance that maintain AI companies' incentives to mitigate risks via means that are not legible to insurers.

Third, exposure to liability might suppress the production and disclosure of the very information on which AI governance depends. Ramakrishnan has developed this objection with care: a developer that investigates the harms its systems have caused, or the dangerous propensities they harbor, generates evidence that may later be marshaled against it, and a developer that discloses what it finds invites suit.⁷⁹ Because frontier expertise and telemetry are concentrated inside the developers themselves, society relies on those companies to understand frontier risk, and Ramakrishnan argues that acquiring better information about the risks and harms of frontier AI “is the most urgent and important task of AI governance at the technological frontier.”⁸⁰ Because “any form of compensatory liability” carries the effect, and no modification of the liability regime is likely to counteract it adequately, he concludes that ex post mechanisms cannot “serve as the principal institutional strategy” for governing frontier AI.⁸¹

The objection bundles two distinct margins: the incentive to investigate risks before harm occurs, and the incentive to disclose harms after they occur. On the investigation margin, the objection cuts in favor of strict liability. Under a negligence regime, what a developer learns defines what was foreseeable, and the file it creates becomes the plaintiff's evidence of breach. Under strict liability, exposure turns on causation and the categorical dangerousness of the activity rather than on what the defendant's documents show it knew, so pre-deployment testing informs the developer's own training, deployment, and insurance decisions while bearing far less

⁷⁸ See, e.g., BRYAN CAPLAN, *THE MYTH OF THE RATIONAL VOTER: WHY DEMOCRACIES CHOOSE BAD POLICIES* 1–18 (2007); CHRISTOPHER H. ACHEN & LARRY M. BARTELS, *DEMOCRACY FOR REALISTS: WHY ELECTIONS DO NOT PRODUCE RESPONSIVE GOVERNMENT* 1–25 (2016).

⁷⁹ Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 182–83).

⁸⁰ *Id.* (manuscript at 183).

⁸¹ *Id.* (manuscript at 182, 184–85) (“any form of compensatory liability will have an effect of this kind”; “it will become increasingly untenable to maintain that ex post mechanisms can serve as the principal institutional strategy for addressing their dangers”).

on its liability.⁸² Ramakrishnan anticipates this response and deems any such cure “a partial one, at most”: even under strict liability, he observes, culpability evidence invites punitive damages and inflated pain-and-suffering awards, and because strict liability pays on causation alone, it magnifies the disvalue of creating causation evidence.⁸³ The culpability-evidence concern turns on the punitive-damages trigger, not on whether governance should rely primarily on liability. The prevailing trigger, conscious disregard of a known risk, does discourage developers from investigating potential risks. The trigger proposed in prior work, and incorporated in this Article’s framework⁸⁴—generating uninsurable risk, evidenced by a near-miss—attaches to the risk itself rather than to the developer’s records, so testing can reduce exposure by helping the developer find and mitigate that risk. Disclosure works similarly: what a developer shares with insurers or regulators under the confidentiality protections the framework specifies is inadmissible in later punitive-damages proceedings, and underwriters can refuse coverage for inadequate transparency while rewarding developers that submit to external audits and share the results with more favorable premiums. Nor must insurers take candor on faith: the framework provides for employee-level evaluator access, internal-test documentation duties, and whistleblower protections. Disclosure then reduces liability exposure only to the extent that it actually reduces risk, the incentive a well-designed regime should create. The causation-evidence concern overstates what testing records contribute under strict liability, where causation is

⁸² Gabriel Weil, *Abnormally Dangerous Algorithms: The Case for Strict Liability at the AI Frontier* 16 (2026) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6588958 [hereinafter Weil, *Abnormally Dangerous Algorithms*] (the rewards-ignorance analysis: foreseeability suffuses breach adjudication, so the developer that declines to investigate can plead unforeseeability while the developer that tests generates the documents on which a breach finding is built; foreseeability persists in strict liability’s scope-of-liability element, but categorically—the question is whether the harm is of the kind that makes the activity dangerous, not what the particular developer knew or recorded).

⁸³ Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 147–48) (culpability evidence “is likely in practice to substantially increase its financial exposure, by making available punitive damages” and inflating pain-and-suffering awards; “a strict-liability regime magnifies the expected disvalue to a company of acquiring or creating evidence that its products are capable of, or likely to, cause” harm) (citing Wendy E. Wagner, *Choosing Ignorance in the Manufacture of Toxic Products*, 82 CORNELL L. REV. 773 (1997)).

⁸⁴ See Weil, *supra* note 10, at 45–49 (proposing elimination of the malice-or-recklessness predicate in favor of a predicate conditioned on uninsurable risk evidenced by practically compensable near-miss harm); Gabriel Weil, *Overcoming Judgment-Proofness: The Law & Economics of Insuring and Mitigating AI Risk* 29 (2026) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6173619 [hereinafter Weil, *Overcoming Judgment-Proofness*] (materials shared with insurers or regulators under protective order are “not, by itself, an admission of liability” and “inadmissible in later punitive-damages proceedings except to impeach inconsistent statements,” on the model of Fed. R. Evid. 408 and 410, a design that “promotes candid disclosure while avoiding perverse penalties for cooperation”); *id.* at 16 (mandated employee-level evaluator access, internal-test documentation duties, and whistleblower protections as the anti-gaming apparatus); *id.* at 25 (incident reports structured to preserve privilege, remain inadmissible in civil actions, and defeat non-disclosure agreements).

primarily established by the incident and its forensics. If anything, such records bear more heavily on negligence's own elements: they show what the developer knew and could have done, and they supply the counterfactual that negligence causation requires, that a specific forgone precaution would have prevented this injury.⁸⁵ This concern does bite where forensics run thin: contested attribution, open-weight releases, harms mediated by other actors, and novel harm mechanisms whose general causation is contested. But the framework proposed here does not leave the evidence base in those cases to litigation incentives; its traceability requirements mandate transparency, information preservation, and incident reporting *ex ante*.⁸⁶ Nor can a developer escape near-miss punitive damages by refusing to look. Under prevailing doctrine, a developer that avoids learning about a risk can rarely be shown to have consciously disregarded it, so ignorance lowers punitive exposure.⁸⁷ Under my proposal, punitive damages would be based on the expected value of the uninsurable risks generated, so anything that mitigates risk would also reduce liability exposure. A plaintiff can prove the predicate with the near-miss event itself, with evaluations conducted under the framework's access mandates, and with the protocol anchors the companion paper specifies; internal records add little to that showing.⁸⁸ Testing that reduces expected risk lowers liability exposure under either regime: under strict liability, because less risk means less expected harm and lower premiums; under negligence, because such testing is part of reasonable care.⁸⁹ Testing that does not reduce expected risk should earn nothing under a correctly applied negligence standard either. So negligence only differentially favors testing to the extent that factfinders err, crediting diligence that could not be expected to offer cost-effective risk mitigation. Nor are the likely errors symmetric. A record identifying the very risk that materialized tells the factfinder the developer knew, and hindsight makes the harm look preventable: mock jurors judging the same precaution decision find negligence far more often once told the harm occurred, and they punish companies for having conducted formal risk

⁸⁵ See Mark F. Grady, *Untaken Precautions*, 18 J. LEGAL STUD. 139 (1989) (negligence liability requires that the identified untaken precaution would have prevented the plaintiff's injury).

⁸⁶ See *infra* Section III.A.

⁸⁷ Cf. *Global-Tech Appliances, Inc. v. SEB S.A.*, 563 U.S. 754, 769 (2011) (willful blindness requires a subjective belief in a high probability of the fact and deliberate action to avoid learning it); Wagner, *supra* note 83 (documenting strategic ignorance in toxic-substances litigation).

⁸⁸ Weil, *Overcoming Judgment-Proofness*, *supra* note 84, at 13–16, 47–53 (evaluator-access mandates; test protocols, validation, and documented error rates as the admissible anchors for the near-miss showing).

⁸⁹ Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 147) (acknowledging that a company that investigates and discloses “might invoke its diligence as proof of reasonable care”).

analyses at all.⁹⁰ Under strict liability the same record sits at the edge of the case, bearing only on factual cause and scope of liability. His related concern that strict liability would mute plaintiff-side investigation, the breach inquiry's discovery function,⁹¹ runs up against volume: a regime that lets meritorious claims proceed without a breach showing puts more claims, and more defendants' files, into discovery, and the disclosure duties both sides of this debate endorse produce the same information directly rather than as a byproduct of litigation.

The disclosure margin is different, and here the objection identifies a real externality that no liability standard dissolves: a developer that already knows of a harm has, other things equal, no tort-law reason to tell anyone. But consider the remedies Ramakrishnan himself canvasses: supracompensatory damages for defendants who failed to discover and disclose the incidents underlying a successful suit; supracompensatory damages for defendants who failed to retain credibly independent evaluators; personal liability for officers and directors who dispensed with independent evaluation before a catastrophe; and penalties measured against precautions a comparably situated developer adopted.⁹² Each of these is an ex post accountability instrument, a damages multiplier or a liability rule, triggered by adjudicated harm and calibrated to the defendant's information practices. They are, in this Article's terms, refinements of the liability regime's parameters, and they sit comfortably beside the near-miss punitive damages framework this Article builds on, which similarly conditions supracompensatory recovery on what the defendant's conduct revealed about undisclosed risk.⁹³ The cure on offer for liability's information pathology is not regulation in place of liability; it is liability, differently parameterized, supplemented by disclosure duties.

That supplement is where the disagreement, examined closely, runs out. The regulatory program Ramakrishnan endorses arrives as transparency and reporting requirements: the entity-

⁹⁰ Kim A. Kamin & Jeffrey J. Rachlinski, *Ex Post ≠ Ex Ante: Determining Liability in Hindsight*, 19 LAW & HUM. BEHAV. 89 (1995) (participants evaluating an identical flood-precaution decision judged it culpable far more often in hindsight than in foresight); W. Kip Viscusi, *Corporate Risk Analysis: A Reckless Act?*, 52 STAN. L. REV. 547 (2000) (mock jurors punished companies for performing risk analyses, and even sound benefit-cost analyses did not reduce the likelihood of punitive awards); cf. Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 146–47) (predicting that juries confronting egregious harm will not credit developer claims of reasonable care).

⁹¹ Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 141) (arguing that fault-based litigation gives plaintiffs incentives to “investigate and publicize the risk-management procedures . . . and risk-benefit tradeoffs” of AI developers, with the automotive-recall literature).

⁹² Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 184).

⁹³ See *supra* Section II.A; Weil, *supra* note 10, at 45–49 (near-miss punitive damages conditioned on uninsurable risk).

based framework he has developed with Dean Ball assigns frontier regulation “the chief task, at least for now,” of “enabling policymakers and the public to better understand the nature and magnitude of the risks posed by cutting-edge AI development.”⁹⁴ Information-forcing rules of that kind are what Part III.A of this Article classifies as supportive policies, traceability requirements that make a liability regime workable, and this Article endorses them on exactly that ground.⁹⁵ Ramakrishnan, for his part, does not propose to discard negligence; the regime at the center of his framework retains the investigation disincentives he himself describes. What remains in dispute is not whether such rules should exist but which instrument anchors the regime they serve.

Fourth, the worst AI harms could exceed any developer’s ability to pay, and liability that ends at bankruptcy cannot price them. This is the judgment-proofness problem.⁹⁶ Ramakrishnan adds a second effect: a developer facing a ruinous judgment may find it rational to gamble for resurrection, racing harder while the case is pending, because insolvency caps what it can lose.⁹⁷ For a frontier developer in financial distress, the socially catastrophic gamble and the privately rational one may coincide.

First, insolvency truncates the incentives of every liability regime, and of every ex post instrument including the administrative fines that regulatory proposals contemplate; it is an argument about the limits of ex post governance generally, not a comparative argument for regulation over liability, unless regulation’s ex ante mandates can supply what no ex post instrument can. Whether they can depends on the regulator knowing which mandates buy safety, the epistemic burden Section A described. Second, the framework this Article defends was built for this problem, and I address it at length in companion work: liability insurance requirements that scale with dangerous capabilities raise the floor of collectable harm to policy limits and put

⁹⁴ Dean W. Ball & Ketan Ramakrishnan, *Entity-Based Regulation in Frontier AI Governance*, CARNEGIE ENDOWMENT FOR INT’L PEACE (July 7, 2025), <https://carnegieendowment.org/research/2025/07/artificial-intelligence-regulation-united-states> (“One of the chief tasks of a frontier AI regulatory regime—arguably the chief task, at least for now—is to put society in a position to reduce such uncertainty and disagreement by enabling policymakers and the public to better understand the nature and magnitude of the risks posed by cutting-edge AI development.”).

⁹⁵ See *infra* Section III.A.

⁹⁶ Steven Shavell, *The Judgment Proof Problem*, 6 INT’L REV. L. & ECON. 45 (1986); Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 183).

⁹⁷ Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 183–84); see Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, 3 J. FIN. ECON. 305, 334–43 (1976).

an underwriter's pricing and monitoring between the developer and the risk, and punitive damages in near-miss cases relocate the incentive to states of the world where judgments are collectable.⁹⁸ Third, gambling for resurrection attaches to any regime that leaves a ruinous liability hanging over a distressed firm (Ramakrishnan notes the effect persists in qualified form even where a bailout is expected),⁹⁹ and its remedies are solvency instruments: insurance and bonding requirements, and the ordinary machinery of insolvency law, all of which this Article's framework treats as supportive policies rather than rivals.¹⁰⁰

The regulatory program Ramakrishnan contemplates is also entity-based: it addresses the developer as a going concern (its protocols, its disclosures, its solvency) rather than the model artifact or the use case.¹⁰¹ So does a liability regime, which is why the two compose. An entity that must maintain capability-scaled coverage, retain independent evaluators, and answer in damages for the harms its systems cause is regulated in the sense that matters, whether the requirements issue from a statute, an insurance contract, or a judgment.

Fifth, Schwarcz and Wolff argue, from the record of cyber insurance, that insurance cannot do what a liability-centered framework asks of it. Insurers there, facing correlated, data-poor, adaptive risks, did not become the private regulators that proponents imagine: they excluded or sub-limited the losses that mattered most, priced on crude proxies like firm size and sector rather than firm-specific safety, oriented their loss-control services toward reducing their insureds' liability rather than the underlying harms, and induced a compliance theater that substituted documentation for safety.¹⁰² On their account, the same structures await AI insurance, and, because insurer transmission is a principal channel by which liability is supposed to discipline behavior, "liability itself will struggle to operate as a reliable regulatory tool."¹⁰³

The default response of courts and insurers to AI risk may well be inadequate, as they predict; Section III.C's case for substitutionary policies rests on the same expectation.¹⁰⁴ The dispute is where special effort is best spent, and the outcomes they catalog are policy choices this framework declines to make. A capability-scaled liability insurance requirement with specified

⁹⁸ Weil, *Overcoming Judgment-Proofness*, *supra* note 84, at 13–23, 40–54; *see supra* Section II.A.

⁹⁹ Ramakrishnan, *Frontier*, *supra* note 14 (manuscript at 183 n.356).

¹⁰⁰ *See infra* Section III.A.

¹⁰¹ Ball & Ramakrishnan, *supra* note 94.

¹⁰² Schwarcz & Wolff, *supra* note 15 (manuscript at 55) (summarizing the cyber-insurance record canvassed in Parts I–III of their article).

¹⁰³ *Id.* (manuscript at 55).

¹⁰⁴ *See infra* Section III.C.

minimum terms forecloses the exclusion strategy; retentions and coinsurance keep the developer's own capital exposed, a design Schwarcz and Wolff themselves suggest when they note that lawmakers could require self-insured layers;¹⁰⁵ and the tail no policy can cover directly is priced through punitive damages in near-miss cases.¹⁰⁶

Frontier AI expertise, unlike cybersecurity expertise, sits inside the insured firms. No underwriter will out-engineer a frontier lab, and this framework does not ask one to.¹⁰⁷ It asks insurers to make developers pay, through risk-rated premiums, for risks their balance sheets could not otherwise absorb, and to verify the mitigations a developer claims. The insurer adds motivation rather than expertise. Insurers can contract with the same independent auditors that other private AI governance proposals rely on. The only difference is that insurers will demand accuracy, whereas AI developers, as audit customers, would tend to favor laxity.¹⁰⁸ Relying on insurers as audit customers fixes the conflict that discredited the issuer-paid credit-rating agencies and that produces the compliance theater Schwarcz and Wolff document. A verifier that carries the risk pays when it materializes, so it prices with care, monitors continuously, and enforces the fixes it requires. Where insurers retained the risk, they built engineering corps for boilers and nuclear plants and wrote airworthiness into aviation coverage; cyber insurers, having excluded the risks that mattered, had no such stake. Anyone who doubts that insurers can play this role owes an account of why government regulators, with weaker incentives and no capital at stake, will do better.

Schwarcz and Wolff's affirmative recommendation is a transparency regime: standardized identification and public disclosure of AI safety risks and mitigation measures, on the model of the loss-data infrastructure that matured insurance markets in domains like workers' compensation.¹⁰⁹ The framework already carries that infrastructure: the companion paper's

¹⁰⁵ Schwarcz & Wolff, *supra* note 15 (manuscript at 56) (noting that lawmakers could "effectively require self-insurance for certain forms of AI liability by prohibiting insurance or mandating large deductibles or co-insurance").

¹⁰⁶ Schwarcz & Wolff, *supra* note 15 (manuscript at 55) (treating bars on insuring punitive damages as deterrence-preserving). On the insurability of punitive damages, see Tom Baker, *Reconsidering Insurance for Punitive Damages*, 1998 WIS. L. REV. 101.

¹⁰⁷ See *supra* note 39 and accompanying text.

¹⁰⁸ Weil, *Overcoming Judgment-Proofness*, *supra* note 84, at 31–39 (the insurer's direct-oversight role, performed in-house or through retained evaluators); Gabriel Weil, *Don't Let AI Developers Hire Their Own Referees*, AI FRONTIERS (July 29, 2026), <https://ai-frontiers.org/articles/dont-let-ai-developers-hire-their-own-referees> (arguing that verification is reliable when the verifier bears the risk it assesses, and that developer-paid verification reproduces the issuer-pays conflict that discredited the credit-rating agencies); Omri Ben-Shahar & Kyle D. Logue, *Outsourcing Regulation: How Insurance Reduces Moral Hazard*, 111 MICH. L. REV. 197 (2012).

¹⁰⁹ Schwarcz & Wolff, *supra* note 15 (manuscript at 55, 57–59) (Part IV.B).

incident-reporting design (reports structured to preserve privilege, remain inadmissible in civil actions, and defeat nondisclosure agreements) is built to deliver “the market-wide learning that liability-tethered insurer processes often fail to produce.”¹¹⁰ Their recommendation is also the destination at which Ramakrishnan’s conclusion and the Ball–Ramakrishnan framework arrive, and it is the category this Article’s Part III.A develops as supportive policy. The leading skeptical treatments of a liability-centered regime each end by recommending information-forcing infrastructure that makes liability and insurance work better. What divides the positions is primacy; all of them require liability to be priced and informed.

Finally, my discussion in Section A may underestimate the practical difficulty of implementing an effective punitive damages regime that accounts for uninsurable risk. Perhaps estimating the uninsurable risks generated by training and deploying specific AI systems will prove intractable, even when we can learn from the practically compensable harm those systems have caused. This objection need not depend on a claim that such estimates will never be feasible, only that they will not perform tolerably well on the timescale that is likely to be relevant for influencing the key AI risk decisions. One potential response to this objection, as suggested above, is that a similar critique applies to any strong AI regulation, which must be based on an implicit estimate of the risks presented by AI systems, either in general or with greater specificity. But proponents of prescriptive regulations or performance standard-based approval regulation might respond that lots of existing regulatory regimes are much more stringent than could be justified by explicit quantitative risk estimates.¹¹¹ Examples include regulatory regimes for nuclear power plants in the U.S. and other countries and the Food & Drug Administration (FDA) approval process for new pharmaceutical products. And it does seem likely that nuclear power and new pharmaceutical products are regulated more stringently, reducing direct risks from those technologies more than a regime of strict liability, punitive damages and liability insurance requirements would. Of course, we can reasonably question whether this risk reduction has come at too high of a cost, in more expensive, dirtier, and more lethal forms of energy generation, and many patients suffering or dying while waiting for the

¹¹⁰ Weil, *Overcoming Judgment-Proofness*, *supra* note 84, at 25.

¹¹¹ See, e.g., CASS R. SUNSTEIN, *LAW OF FEAR: BEYOND THE PRECAUTIONARY PRINCIPLE* 127–30 (2005) (discussing how “probability neglect” and other cognitive biases lead to stringent regulations, such as those for new drugs, that are disconnected from purely quantitative risk assessments).

FDA to approve a drug that might help them.¹¹² Nonetheless, it does seem plausible that we could end up with an AI-governance regime based on prescriptive regulations and approval procedures that reduces AI risk from sources under its jurisdiction more than a liability regime would. This concern is sufficiently distinct and complex that I will address it in its own section.

C. Is Expected Welfare Maximization the Wrong Policy Goal?

Assuming it is politically feasible, the desirability of a governance framework based on extremely stringent approval regulation hinges on three questions. First, given the massive potential downside risks of AI, does such a seemingly hyper-cautious regime actually come out ahead of a well-designed liability regime on ordinary risk-utility analysis? Second, given the potentially catastrophic and uncertain nature of the threats posed by AI, should we abandon ordinary expected utility maximization in favor of a more precautionary approach? Third, do approval regulation or other ex ante prescriptive approaches to AI governance offer substantially greater protections against the most extreme potential AI catastrophes?

Based on the considerations surveyed above, the answer to the first question would seem to be no. If you are convinced that liability offers a more favorable risk mitigation-to-innovation inhibition ratio and other costs than prescriptive and approval regulation, then there is little reason to think the latter would perform better in terms of producing greater expected net utility. To believe that, you would have to think that a liability-centered AI governance regime would systematically tend to either over- or underdeter the risk-generating activities.

It is not plausible to conclude that the implicit risk assessments embedded in prescriptive regulations and approval systems will tend to be more accurate than explicit estimates used to calculate punitive damages, which can be conducted in a superior epistemic position. Perhaps judges and juries, relying on expert testimony and other evidence presented in an adversarial process, are poorly positioned to estimate the risks posed by AI systems. But the case for *liability* as the centerpiece of AI governance does not entail the conclusion that *courts* are the best institutional forum for imposing liability. Perhaps it makes sense for fact-finding, at least in cases

¹¹² See, e.g., Wade Allison, *The Bad Science Behind Expensive Nuclear*, WORKS IN PROGRESS, June 12, 2024, <https://worksinprogress.co/issue/the-bad-science-behind-expensive-nuclear/> (arguing that the ALARA standard, based on the scientifically disputed Linear No-Threshold model, imposes excessive and unnecessary costs on the nuclear industry for infinitesimal benefits, thereby making clean energy unaffordable); Daniel B. Klein, *Drug-Approval Denationalization*, LIBR. OF ECON. & LIBERTY (Apr. 6, 2009), <https://www.econlib.org/library/Columns/y2009/Kleindrugapproval.html> (arguing that pre-market approval imposes extreme costs, delays, and uncertainties that “suppress the development of new therapies” and that this regulatory system prevents many life-saving drugs from ever being created).

involving punitive damages designed to mitigate uninsurable risks, to happen before an executive branch agency staffed with AI risk experts.¹¹³ Perhaps it even makes sense to use civil fines paid to the government instead of punitive damages paid to private plaintiffs to address uninsurable risks. Whether these changes are desirable is beyond the scope of this Article, and is distinct from the choice between ex post liability and ex ante regulation as the primary AI governance mechanism.

Even if an extremely stringent approval regime fails an ordinary risk-utility test, might it still be desirable if a precautionary approach is warranted and such a regime offers substantially greater protections against the most extreme potential AI catastrophes? One candidate for a precautionary approach is the maximin rule: choose the option with the best worst-case outcome. As Cass Sunstein explains, weaker forms of maximin are just reminders to take catastrophic risk seriously and avoid the temptation to neglect low-probability, high-impact events and are consistent with ordinary expected value maximization.¹¹⁴

This reminder may be particularly important with regard to the treatment of existential risks by our tort liability system. That is, courts may be reluctant to place adequate negative value on actions that risked not only the lives of presently living humans and their foreseeable descendants, but the entire future of human civilization. While there are good reasons to be skeptical that punitive damages would, in practice, place adequate value on the interest of hypothetical future people, this does not bear on the design features of an *ideal* AI governance regime. An ideal liability regime could assign appropriately high value to preventing existential catastrophe. In any case, similar practical constraints are likely to come to bear on implicit valuations of future people in regulatory policies — politicians and voting publics simply place little weight on the value of the long-term future. Thus, the normative imperative to adequately value the future does not clearly favor ex ante mechanisms over liability.

But stronger forms of maximin would counsel caution even when the precautionary approach does not maximize expected utility. John Rawls, for instance, contends that maximin is justified (a) in the face of potentially catastrophic outcomes, (b) when probabilities cannot be

¹¹³ Full adjudication of tort suits before an executive agency is likely unconstitutional under *SEC v. Jarkesy*, 603 U.S. 109, 140 (2024) (holding that the Seventh Amendment entitles a defendant to a jury trial when the SEC seeks civil penalties for securities fraud).

¹¹⁴ Cass R. Sunstein, *Irreversible and Catastrophic*, 91 CORNELL L. REV. 841, 870-79 (2006), <https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=3029&context=clr>.

assigned, and (c) when the loss from following maximin is a matter of relative indifference. Stephen Gardiner adds a fourth criterion: (d) that the potentially catastrophic must satisfy some minimal threshold of plausibility.¹¹⁵ As Sunstein points out, unless we relax condition (c), this form of maximin does not actually counsel precaution in many real-world situations; eliminating risks generally comes with substantial costs.¹¹⁶ In particular, any conceivable strategy for eliminating the possibility of the worst outcome is likely to require enormous sacrifices.¹¹⁷ But Sunstein allows that it may be sensible to abandon (c) in the face of a plausible, but unquantifiable, threat of an irreversible catastrophe. The standard economist's objection to this stronger form maximin is that it implies infinite risk aversion.¹¹⁸ That is, in placing absolute priority on avoiding the worst-case outcome, maximin is willing to trade arbitrarily large gains in non-catastrophic scenarios for incremental reductions in the severity of the worst plausible catastrophe. This objection is typically paired with an argument that infinite risk aversion is neither an accurate description of people's preferences nor a normative ideal suitable to guide policymaking.¹¹⁹

This argument appears convincing as far as it goes, but Sunstein argues that infinite risk aversion is not required to justify maximin under conditions of genuinely unquantifiable uncertainty.¹²⁰ If no probability can be assigned to the worst-case outcome, then we cannot say it requires infinite risk aversion to be willing to pay substantial costs to eliminate it. Frank Knight drew the distinction between quantifiable risk and unquantifiable uncertainty in the early 1920s.¹²¹ Milton Friedman, following I.J. Savage, denied that the latter category exists, insisting that subjective probabilities can always be inferred from individual choices. Sunstein finds this objection unconvincing, arguing that *objective* probabilities are what is needed to make expected utility theory work for catastrophic risks.¹²² John Kay and Mervyn King go further, arguing both

¹¹⁵ See JOHN RAWLS, A THEORY OF JUSTICE 152–57 (1971); Stephen Gardiner, *A Core Precautionary Principle*, 14 J. POL. PHIL. 33 (2006), <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9760.2006.00237.x>.

¹¹⁶ Sunstein, *supra* note 114, at 881–2.

¹¹⁷ Eliminating all risk of fatal car accidents, for example, would require banning automobiles.

¹¹⁸ See Kenneth J. Arrow, *Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice*, 70 J. PHIL. 245 (1973); John C. Harsanyi, *Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory*, 69 AM. POL. SCI. REV. 594, 595 (1975).

¹¹⁹ R. A. Musgrave, *Maximin, Uncertainty, and the Leisure Trade-Off*, 88 Q. J. ECON. 625, 627 (1974).

¹²⁰ Sunstein, *supra* note 114, at 882.

¹²¹ See generally FRANK H. KNIGHT, RISK, UNCERTAINTY AND PROFIT (1921).

¹²² Sunstein, *supra* note 114, at 882–5.

that Friedman misread the limits Savage himself placed on his framework and that what they call *radical uncertainty* is pervasive in human and institutional decision making.¹²³

In the AI context, many observers are willing to offer quantitative estimates for the most extreme negative outcomes, commonly shorthand as $p(\text{doom})$, but these estimates range from a chance of human extinction greater than 99% to well below 1%.¹²⁴ If Knightian uncertainty turns only on the absence of objective probability information, as Sunstein's account suggests, AI risk qualifies. But as Kay and King point out, objective probabilities of the sort inferred from a stable frequentist distribution are a feature of "small-world" models that are rarely applicable to real-world decisions.¹²⁵ In typical policy problems, there is both readily quantifiable risk within the model being used to analyze the problem and uncertainty, resistant to quantification, about how much confidence the model deserves. Kay and King embrace the strong implication that probability theory has limited application to most real-world decisions.¹²⁶ But this may prove too much. As Sunstein points out, there are typically risk-risk tradeoffs. We cannot coherently adopt a policy of acting in a precautionary fashion whenever we face uncertainty, if uncertainty is pervasive.¹²⁷ The sort of uncertainty that can justify invoking maximin must be more narrowly circumscribed. The dangers posed by AI plausibly fall into that narrower category. It is not just that confidence in any particular model resists quantification; the development of strong AI is a truly unprecedented event for which it is difficult to define an appropriate reference class.¹²⁸ Another wrinkle is that those least willing to offer quantitative probability estimates for an AI catastrophe tend to be those skeptical of AI risk, not those most worried.¹²⁹ Thus, counterintuitively, the epistemic approach favored by the risk skeptics is more concordant with the usual justification for precautionary policies, difficulty assigning probabilities. In any case,

¹²³ See generally JOHN KAY & MERVYN KING, *RADICAL UNCERTAINTY: DECISION-MAKING BEYOND THE NUMBERS* (2020).

¹²⁴ There are also people who claim with absolute certainty that AI will not produce an existential catastrophe, but these statements are not best interpreted as probability estimates.

¹²⁵ See KAY & KING, *supra* note 123, at 37–41 (distinguishing between "small worlds," where probabilities are known, and the "large world" of radical uncertainty that characterizes most real-world decisions).

¹²⁶ *Id.* at 72–76.

¹²⁷ As an example within the domain of AI policy, extreme upsides of developing advanced AI could include inventions that reduce other catastrophic risks, such as through the invention of better technology to prevent pandemics. Thus, it is not obvious whether a precautionary approach counsels over- or under-detering negative externalities from AI development.

¹²⁸ On the importance of choosing an appropriate reference class when forecasting risk, see generally PHILIP E. TETLOCK & DAN GARDNER, *SUPERFORECASTING: THE ART AND SCIENCE OF PREDICTION* (2015).

¹²⁹ See Arvind Narayanan & Sayash Kapoor, *AI existential risk probabilities are too unreliable to inform policy*, AI SNAKE OIL (July 26, 2024), <https://www.aisnakeoil.com/p/ai-existential-risk-probabilities>.

those offering such probability estimates appear to treat them as subjective Bayesian probabilities, not the sort of objective probabilities that Sunstein thinks expected value calculations require.

In the later work, Sunstein offers some further gloss on the conditions under which maximin is warranted, in the form of proposed language for OMB Circular A-4, which offers federal agencies guidance on cost-benefit analysis.

In special circumstances, you might consider avoiding the worst-case scenario and thus following the maximin rule, which calls for eliminating the worst of the worst-cases. The strongest cases for following that rule would involve three factors: (1) Knightian uncertainty, understood as an inability to assign probabilities to various options; (2) catastrophic or grave consequences from one option, but not from other options; and (3) low or relatively low costs, or low or relatively low benefits foregone, as a result of choosing the option that avoids the worst-case scenario. Again in cases of uncertainty, more difficult cases, in which (for example) the costs of avoiding the worst-case scenario are very high, might be subjected to use of the maximin principle if (for example) the worst-case scenario is genuinely catastrophic.¹³⁰

Application of this approach to AI risk is far from straightforward. Assume criterion (1) is satisfied. Under criterion (2), inadequate approaches to AI regulation could result in catastrophic or grave consequences, but so could the sort of draconian limits on AI research, development, and deployment that might be needed to eliminate those potential outcomes. That is, the sort of tight global controls on AI that would likely be part of any effort to eliminate the possibility of the worst potential harm would risk global dictatorship or world war, and likely leave us more vulnerable to other disasters. Similarly, for criterion (3), a maximin approach to the threat from AI, if one is feasible, would likely forgo enormous potential economic benefits from AI advancements. Nonetheless, the final sentence of the quoted language reopens the door to applying maximin to AI in “more difficult cases,” where the costs of avoiding the worst-case scenario are very high but the scenario is “genuinely catastrophic.” The worst-case AI scenarios, which include outright human extinction, surely qualify.¹³¹ But here Sunstein’s guidance seems to run out. “In the hardest and most intriguing cases, it is not possible to defend any simple rule.

¹³⁰ CASS R. SUNSTEIN, *AVERTING CATASTROPHE: DECISION THEORY FOR COVID-19, CLIMATE CHANGE, AND POTENTIAL DISASTERS OF ALL KINDS* [pin] (2021).

¹³¹ Some scenarios, often addressed under the banner of suffering risk or S-risks, are arguably worse than extinction, and involve the permanent enslavement of the remaining human population under intolerable conditions.

Some kind of judgment must be made. Nothing in decision theory can specify that judgment.”¹³² What are we to do with a problem where there is unquantifiable uncertainty about the likelihood of a genuinely catastrophic outcome, but the costs of preventing it with high confidence are also enormous and include catastrophic worst-case scenarios of their own? In his 2006 paper, Sunstein appears to reject maximin under these circumstances: “It makes no sense to take steps to avert catastrophe if those very steps would create catastrophic risks of their own.”¹³³ But it seems plausible, as a matter of judgment, that we might conclude we should act to exclude the first set of risks even if some catastrophic risk remains, based on a qualitative judgment that the former risks are graver or more likely.

Finally, let us turn to the question of whether approval regulation or other ex ante prescriptive approaches to AI governance offer substantially greater protections against the most extreme potential AI catastrophes. There may be no reliable means of eliminating the risks of catastrophic AI misalignment or misuse, even if we are willing to tolerate the creation of alternative potential catastrophes in the process. A permanent ban on research, development, and deployment of AI systems that can be expected to have potentially dangerous capabilities would be an enormous political lift. Even if it could be enacted, it may not be stable in the medium to long term. Even if we ignore domestic political considerations, as may be appropriate when considering ideal policy, there is no clear pathway to global enforcement of such a ban.¹³⁴ If the U.S. were to unilaterally cease AI development, for instance, others would likely catch up to the leading U.S. labs within a few years. And if the U.S. were to persuade its Western allies to shut down potentially dangerous AI development and tighten export restrictions on cutting-edge computer chips to China, this would likely only delay global AI progress by a decade or two at most, and potentially leave the technology in the hands of less cautious actors. Military action against Chinese data centers and other elements of their AI infrastructure could extend this timeline somewhat, but likely not indefinitely, and would generate its own catastrophic risks, such as nuclear war. Over time, algorithmic progress is likely to decrease the quantity of

¹³² Cass R. Sunstein, *Knightian Uncertainty in the Regulatory Context*, 9 BEHAV. PUB. POL’Y 614, 628 (2025), <https://doi.org/10.1017/bpp.2024.59>.

¹³³ Sunstein, *Irreversible and Catastrophic*, *supra* note 114, at 893.

¹³⁴ See, e.g., TIM FIST, UPHILL BATTLES: THE VERIFICATION CHALLENGES OF ARTIFICIAL INTELLIGENCE ARMS CONTROL 3–5 (Center for a New American Security, 2020) (detailing the profound technical and institutional challenges in monitoring and verifying compliance with any potential AI arms control regime, including the dual-use nature of AI components, the lack of reliable monitoring techniques, and the difficulty of attributing illicit activity).

compute needed to train powerful models, and chip infrastructure may also get easier to hide. Conversely, less restrictive U.S. AI regulatory regimes would likely fail to fully eliminate the possibility of domestic AI catastrophe. Even if strong global coordination were achieved, either across national governments or via consolidation of the global population under a single political unit, a regime with the power to enforce a total permanent ban on the development of potentially dangerous AI systems would have to be powerful and unchecked enough to engage in abuses of its own. Even if we rule out non-AI threats, there might be little to prevent the leaders of such an unchecked regime from internally developing potentially dangerous AI.

One plausible approach to circumstances in which we cannot eliminate the possibility of existential catastrophe is to pursue Nick Bostrom's notion of maxipok: maximizing the probability of an OK outcome. In Bostrom's analysis, the motivation for maxipok is the large discontinuity in value between an existential catastrophe and a near-existential catastrophe. Whereas conventional cost-benefit analysis and common sense would treat the difference between peace and war that kills 99% of the human population as much larger than the difference between a war that kills 99% of the human population and a war that kills 100% of humans, Bostrom — following Derek Parfit — considers the latter difference to be much larger due to the loss of all potential future value of human civilization.¹³⁵

While maxipok has some appeal, two points will make clear that Bostrom's analysis has little bearing on the choice between ex post liability and ex ante regulation. First, in order to implement maxipok, we must have probability estimates for existential catastrophe under various policies, or at least an ordinal ranking of them. But if we have such information, we are probably not facing genuinely unquantifiable uncertainty, such that a departure from ordinary expected utility maximization is warranted. Bostrom would not dispute this; he just thinks the probability of existential catastrophe dominates the expected utility calculation. This leads to the second point. If the value of human extinction is genuinely orders of magnitude worse than near-extinction, that can be accounted for in the punitive damages calculation for a liability regime. For instance, legislation could be enacted stipulating a dollar value for human extinction, so the jury would only be tasked with assessing the ex ante likelihood that the defendant's conduct would have resulted in human extinction. There may be political barriers to putting the sort of

¹³⁵ Nick Bostrom, *Existential Risk Prevention as Global Priority*, 4 GLOB. POL'Y 15, 17-9 (2013), <https://existential-risk.com/concept.pdf>.

weight on the interest of future generations that Bostrom’s analysis entails, but those barriers would tend to apply to any law or policy instrument designed to protect against existential risks, including highly restrictive approval regulation.¹³⁶ It’s not clear why we should expect courts, as a rule, to place less weight on the interests of future generations than regulatory agencies, especially if they are instructed to do so by legislation.

III. The Role of Other Instruments in a Liability-Centered AI Governance Regime

The foregoing analysis has emphasized two key premises. First, liability has clear limits on its reach as a policy tool. There are some AI policy problems it cannot be called upon to solve. Second, with circumscribed exceptions like cybersecurity for frontier systems, it is the best governance tool within its wide domain of applicability, at least as a matter of ideal policy. That is, liability should be the centerpiece of AI governance, but leaves important gaps. This Part explores the implications of these premises for other AI governance tools. In doing so, it identifies three channels by which other AI law and policy instruments can contribute to a liability-centered governance regime. Section A discusses the role that *supportive* policies can play in facilitating the implementation of a strong liability regime. Section B examines how *complementary* policies can fill the gaps left by even a highly robust liability regime. Finally, Section C analyzes how *substitutionary* policies can help compensate for the shortcomings of an inadequate liability regime like the status quo negligence regime.

A. Supportive Policies

The current liability regime applicable to AI harm in the U.S. is inadequate along multiple dimensions. The most important of these deficiencies can be addressed by changes in liability doctrine as such, via either judicial rulings or legislation. These changes include adopting strict liability for the harm arising from the inherently dangerous nature of training and deploying frontier AI systems; assessing proximate cause by asking whether the prospect of liability would have offered the defendant meaningful incentives to mitigate the risk of the harm for which they are sued; and applying punitive damages in “near-miss” cases of practically compensable harm associated with uninsurable risks, without a requirement that the defendant

¹³⁶ See Daniel Carpenter, Approval Regulation for Frontier Artificial Intelligence: Pitfalls, Plausibility, Optionality, 3 (2024) available at https://lwp.law.harvard.edu/sites/projects.iq.harvard.edu/files/pegroup/files/carpenter_approval-regulation-for-frontier-ai-regulation-20240111.pdf.

has acted recklessly or with malice. But these changes would not fix all of the shortcomings of liability, even within its domain of core competency.

For instance, it may be difficult in some cases to attribute harm suffered by a given plaintiff to particular AI systems and particular upstream human conduct. Requirements for transparency, information preservation, watermarking, incident reporting, and identity binding could greatly facilitate the ex post attribution of AI harms to the responsible human actors.¹³⁷ Such requirements are likely to impose a modest compliance burden compared to prescription rules or approval regulations designed to directly ensure the safety of AI systems.¹³⁸ Accordingly, they can be justified based on relatively low risk perceptions. Like ex post liability, they do not depend on a robust ex ante consensus that is convinced that AI poses large societal risks. Modest ex ante requirements that facilitate ex post attribution of AI harms are broadly consistent with a liability regime centered on ex post liability.

One obvious gap left by this framework is that, by default, not all legally compensable harm that arises from insurable risks will be practically compensable. That is, some defendants may be practically judgment-proof even for large compensatory or punitive damages awards, even if they could theoretically have purchased insurance against such a liability judgment. Indeed, with regard to the outer limits of insurable risks, AI developers have weak incentives to ensure their ability to pay for the liabilities they might incur. If their systems fail catastrophically or fail non-catastrophically in a manner associated with uninsurable risk, this is likely to deal a substantial blow to the value of their enterprise even if they do carry adequate liability insurance. Accordingly, unless they are required to demonstrate their ability to pay for harm their systems might cause, AI developers may be able to externalize a substantial fraction of the *insurable* risks generated by their system, even if they expect to be held legally liable for them.

Fortunately, liability insurance or bonding requirements can fill this gap. Developers of frontier AI systems, as defined by statute, could be required to submit their models for pre-deployment testing. If that testing reveals potentially dangerous capabilities, a state or federal regulatory agency could set a minimum liability coverage requirement, which could be satisfied by securing an insurance policy or setting aside sufficient capital to pay out the required amount.

¹³⁷ Alan Chan et al., *Infrastructure for AI Agents 3*, arXiv.org (2025), <https://arxiv.org/abs/2501.10114> (last visited Aug. 4, 2026). California has enacted first-generation requirements along these lines. *See* Transparency in Frontier Artificial Intelligence Act, S.B. 53, 2025–2026 Leg., Reg. Sess. (Cal. 2025).

¹³⁸

If the model developer elects to submit a binding deployment plan, with access to the model subject to specified restrictions, the regulator could account for these specifications in determining the coverage requirement.¹³⁹ For example, releasing a model with open weights may require more liability coverage than limiting access to an API, combined with a robust cybersecurity regime. If the regulator determines that the appropriate coverage requirement is larger than the maximum feasible insurance policy, then the coverage requirement would be set at that determined maximum value. This maximum insurable value would also serve as a benchmark for courts to use in assessing punitive damages. Any risks of harm smaller than the maximum insurable value could be handled with ex post compensatory damages in cases when those risks are realized. But risks of harm greater than the maximum insurable value can only be addressed indirectly, using punitive damages in cases of practically compensable harm that are associated with the uninsurable risk.

In principle, the regulator that sets the liability insurance requirement could also be tasked with assessing whether the prospect of liability, including punitive damages, can adequately mitigate the risks posed by the system under evaluation. There are two related reasons why that condition might not obtain. First, the regulator may have reason to believe that the sort of uninsurable risks posed by the system are unlikely to generate associated “near-miss” or “warning shot” cases. As discussed in the introduction, punitive damages are only an effective tool for mitigating uninsurable risk if the developers and deployers of AI systems, and their insurers or investors, have reason to expect that warning shots are substantially more likely than actual uninsurable catastrophes. If a warning shot is only 10 times more likely than a given uninsurable catastrophe, then punitive damages can only be used to expand the range of risks addressed to 10 times the maximum insurable value. For some extreme risks, this may be inadequate. Second, even if warning shots are likely, the magnitude of the uninsurable risk may simply be too large to address via a practically feasible damages award, even indirectly. To take an extreme example, if a regulator determines that the deployment of a specific AI system would generate a 5% risk of human extinction, then the punitive damages award that would be triggered by a near-miss case would itself be uninsurable. In both such cases, it may make sense to

¹³⁹ In defiance of this logic, the European AI Act reduces regulatory requirements for open access models. “Free and open licence GPAI model providers only need to comply with copyright and publish the training data summary, unless they present a systemic risk.” High-level summary of the AI Act, EU Artificial Intelligence Act, <https://artificialintelligenceact.eu/high-level-summary/> (last visited Aug. 4, 2026).

authorize the regulator to outright deny the developer permission to deploy the model. This would represent a hybrid between ex ante approval regulation and ex post liability, with approval regulation stepping in when liability hits its limit. This approval authority could also be used to prevent models from being deployed without first undergoing pre-deployment testing and securing the required amount of insurance.

A potential failure mode for this proposal hybrid approach is that the regulator might make excessive use of its disapproval authority. The Food and Drug Administration approval regime for new pharmaceutical products and the Nuclear Regulatory Commission approval regime for nuclear power plants are clear examples of the tendency for regulators empowered with disapproval authority to overreach. As a result, the model-disapproval backstop could swallow the deterrence-based liability and liability insurance framework that part II argued we should prefer. Not only would this potential collapse into an approval regime make for suboptimal policy, but the prospect of such collapse may threaten the political viability of the whole framework. Both of these risks could be mitigated by subjecting disapproval decisions to judicial review, with the burden put on the regulator to demonstrate that the model is too dangerous to deploy, even with the maximum feasible liability insurance and a clear expectation of punitive damages for harm associated with uninsurable risk.

This raises the question of how both liability insurance requirements and backstop approval regulation should be enforced. The standard answer to this problem in most regulatory domains is steep civil fines, often backed by the threat of criminal prosecution. But some AI governance researchers have suggested a more reliable enforcement mechanism: compute governance. Under a robust compute governance framework, AI developers would only have access to the computational resources needed to train and deploy their models if they receive regulatory approval.¹⁴⁰ The detailed functioning of, and tradeoffs implicated by, this sort of compute governance are beyond the scope of this Article. But it is worth noting that compute governance is equally available as a potential tool for enforcing a licensing regime that relies primarily on liability insurance as it is for standard approval regulation regimes.

There are also two key technical hurdles to implementing the strong form of liability for which I have advocated. First, regulators will need to develop evaluation tools that enable them

¹⁴⁰ Markus Anderljung et al., *Frontier AI Regulation: Managing Emerging Risks to Public Safety*, arXiv.org (2023), <https://arxiv.org/abs/2307.03718> (last visited Aug. 4, 2026).

to estimate, with reasonable accuracy, the maximum damage that a particular AI system, deployed according to a specified protocol, could cause. Second, in cases where an AI system is responsible for practically compensable harm, we will need methods of estimating the magnitude and relative elasticity¹⁴¹ of any uninsurable risks that were generated by the key system training and deployment decisions. While both government bodies like the U.S. AI Safety Institute and UK AI Security Institute, and nonprofit organizations like Model Evaluation and Threat Research are working on developing methods for evaluating the threats posed by AI systems, they are not yet fit for these purposes. Government investments in developing these tools would help lay the groundwork for implementing a robust liability regime, complete with insurance requirements and punitive damages to account for uninsurable risks.

Both of these projects could also be aided by modest requirements that AI developers register training and deployment plans, self-assess and disclose the risks posed by their systems, and submit their systems to third-party auditors.¹⁴² This will help advance the science of model evaluations and give courts and regulators a window into the risks generated by AI developers' training and deployment decisions. Similarly, watermarking requirements and other policies to facilitate the traceability of AI activities would make it easier to assign legal responsibility for suspected AI harm.¹⁴³ Finally, in jurisdictions where courts and legislatures are unwilling to adopt a strict liability standard for AI harm, some of the benefits of a strong liability could be achieved by specifying duties that spell out or supplement the reasonable-care requirement of negligence law.

B. Complementary Policies

Complementary policies are those best positioned to fill the gaps identified in the introduction of this Article. These gaps include market failures distinct from the core risk externality, including the public goods problem associated with AI alignment and safety research, and the winner-take-all race dynamics that might push AI developers to advance the capabilities frontier too rapidly, even if they are forced to internalize the risks directly generated

¹⁴¹ Weil, *supra* note 10, at 59-60.

¹⁴² Noam Kolt et al., Responsible Reporting for Frontier AI Development, arXiv.org (Apr. 3, 2024), <https://arxiv.org/abs/2404.02675>; Margaret Mitchell et al., Model Cards for Model Reporting, arXiv.org (Jan. 14, 2019), <https://arxiv.org/abs/1810.03993>.

¹⁴³ Siddharth Srinivasan, *Detecting AI fingerprints: A guide to watermarking and beyond*, BROOKINGS INST. (Jan. 4, 2024), <https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/>; John Kirchenbauer et al., A Watermark for Large Language Models, arXiv.org (May 1, 2023), <https://arxiv.org/abs/2301.10226>.

by their training and deployment decisions. Relatedly, liability, like all domestic AI governance measures, is also potentially vulnerable to regulatory arbitrage, absent strong global coordination. Also, some policy challenges are not best understood as market failures at all, like potential misuse of advanced AI systems by government officials. Finally, there are some risk externalities that are unlikely to give rise to enforceable liability judgments, including uninsurable risks for which warning shots are unlikely, actions whose risks may only be realized via highly attenuated or factually murky causal chains, and legally non-compensable harms like political misinformation and election interference.

Like most information goods, breakthroughs in AI alignment and safety are non-rival and not easily excludable. That is, once a discovery is made, it can be shared widely at near-zero marginal cost, and it is hard to charge people who want to access it. Public goods tend to be underprovided by competitive markets, because most of the benefits of their provision accrue to people other than those who bear the costs to produce them. By default, the discoverer of an important AI alignment breakthrough will not reap market rewards commensurate with the value of the discovery. In some domains, policymakers address information-based public goods problem with intellectual property (IP) protections like patents, trade secrets, and copyrights, but IP protections don't work very well for basic research.¹⁴⁴ Even for patentable applied research, patents may not be the best way to encourage more alignment and safety innovation.¹⁴⁵ Patents have the advantage that the incentive naturally scales with the value of the breakthrough, but the disadvantage is that they slow technology diffusion and inhibit downstream innovations that would need to build on patented technologies.¹⁴⁶ Accordingly, it may be preferable to support AI alignment and safety innovation primarily via a combination of direct government research, grants and other subsidies, and prizes for specified breakthroughs.¹⁴⁷ Private philanthropy also has a role to play in both supplementing government subsidies and experimenting with funding models that may not be politically viable until they have a demonstrated track record.¹⁴⁸

¹⁴⁴ Joseph Stiglitz, *Economic Foundations of Intellectual Property Rights*, 57 DUKE L.J. 1693, 1697 (2008), <https://scholarship.law.duke.edu/dlj/vol57/iss6/3/>.

¹⁴⁵ See, e.g. Dennis Crouch, *Using Intellectual Property to Regulate Artificial Intelligence*, 89 MO. L. REV. 781, 788-9 (2024), <https://scholarship.law.missouri.edu/cgi/viewcontent.cgi?article=4679&context=mlr>.

¹⁴⁶ Stiglitz, *supra* note 144, at 1697-1706.

¹⁴⁷ Stiglitz, *supra* note 144, at 1721.

¹⁴⁸ Matthew Erskine, *Innovative Philanthropy: Beyond Traditional Grantmaking*, FORBES (Jan. 30, 2024, 4:43 PM), <https://www.forbes.com/sites/matthewerskine/2024/01/30/innovative-philanthropy-beyond-traditional-grantmaking/>.

Even if AI developers expect to pay for all the harm caused directly by their AI training and deployment activities, this may not fully align their interests with social welfare. If any AI developer expects to reap outsized gain if they are the first to reach some milestone, like artificial general intelligence (AGI), they have incentives to rush and skimp on any precautions that would slow down progress.¹⁴⁹ A similar dynamic may play out on the geopolitical level, where great power competition may drive governments to adopt policies that favor faster, less cautious, domestic AI development in order to win a race with geopolitical rivals. This dynamic seems to play a substantial role in U.S. AI policy, manifesting in export controls designed to hamper Chinese AI development paired with subsidies to support the domestic chip supply chain and a light-touch approach to AI regulation. At both scales, emphasis on the zero-sum conflict on the race to AGI (or another conception of strong AI), results in actions that are suboptimal from a global social welfare perspective.

Policy options for mitigating race dynamics may be limited, especially in the absence of strong institutions to enable global coordination. The emerging U.S.-led export control regime arguably helps tamp down race dynamics, but also comes with substantial risks. These include reinforcing conflict mindset and undermining cooperative approaches to global AI governance, encouraging China to build out a parallel chip manufacturing supply chain, and even potentially helping to provoke a Chinese invasion of Taiwan, where most high-end chips used to train AI models are manufactured.

On the domestic front, efforts to secure model weights and other key proprietary information about frontier AI systems from cybersecurity breaches could help the leading domestic labs keep their leads and feel less pressure to hastily and incautiously advance the capabilities of their models. While both competitive pressures and the risk of liability if their models are misused may give labs strong incentives to secure their models, they may lack the capacity to do so, at least under attack by a determined nation-state. Direct regulation and even direct government provision of cybersecurity services for frontier labs may be warranted.¹⁵⁰ After all, in contrast to the situation with AI technology itself, cybersecurity expertise is

¹⁴⁹ See generally Armstrong et al., *supra* note 21.

¹⁵⁰ Leopold Aschenbrenner, *Lock Down the Labs: Security for AGI*, in *Situational Awareness* (2024), <https://situational-awareness.ai/lock-down-the-labs/>.

concentrated in government agencies like the National Security Agency and the Department of Defense, not the AI labs.¹⁵¹

However, the greatest source of information-based race acceleration may not arise from the leaking of model weights. Instead, acceleration of capabilities progress may be driven more by the publication of basic scientific findings, like Google’s paper on transformers, a key technical breakthrough on which all large language models rely.¹⁵² In recent years, leading AI labs have shifted to releasing less information about their models and any improvements in algorithmic efficiency or other technical advances.¹⁵³ This shift seems to be mostly motivated by business reasons, but likely also reflects some concerns about AI safety. Whatever the motivation, AI developers remain free to disclose technical advances and any governmental effort to restrict such disclosures would face substantial practical, political, and constitutional barriers.¹⁵⁴ In any case, we should not expect liability to be an effective policy lever in this domain. Whether other tools can effectively tamp down race dynamics without excessively burdening free speech or sacrificing too many of the potential benefits of AI advances remains an open question for AI governance researchers.

A potential response to imposition of a strong liability regime in one country or subnational jurisdiction is for investments in AI development and deployment to shift to other jurisdictions. In principle, AI systems could harm U.S. residents without ever being deployed by a person acting in the U.S. For instance, foreign actors might use a foreign AI system to engage in phishing scams to steal the identities of U.S. residents. While legal liability would still apply under domestic law if residents of the jurisdiction that imposed the liability regime are harmed by foreign systems. But if the foreign AI developer lacks substantial assets in the jurisdiction imposing the liability, enforcing any damages award may not be feasible.

¹⁵¹ *Id.*

¹⁵² Ashish Vaswani et al., *Attention Is All You Need*, arXiv.org (2017; last revised Aug. 2, 2023), <https://arxiv.org/abs/1706.03762>. See also Mary Phuong & Marcus Hutter, *Formal Algorithms for Transformers*, arXiv.org (July 19, 2022), <https://arxiv.org/abs/2207.09238>.

¹⁵³ Lauren Castle, *Trade Secrets Summoned to Protect AI Amid Noncompete Uncertainty*, BLOOMBERG LAW (July 16, 2023, 5:03 AM), <https://news.bloomberglaw.com/ip-law/trade-secrets-summoned-to-protect-ai-amid-noncompete-uncertainty>; Memorandum from Quinn Emanuel Urquhart & Sullivan, LLP (Apr. 24, 2024), <https://www.quinnemanuel.com/media/wi2pks2s/the-rising-importance-of-trade-secret-protection-for-ai-related-intellect.pdf>.

¹⁵⁴ Alan Rozenshtein, *There Is No General First Amendment Right to Distribute Machine-Learning Model Weights*, LAWFARE (Apr. 4, 2024, 8:02 AM), <https://www.lawfaremedia.org/article/there-is-no-general-first-amendment-right-to-distribute-machine-learning-model-weights>.

Under the full faith and credit clause of the U.S. Constitution, states are required to enforce the judgments of courts in other states against their citizens. On the international level, however, courts are not under any general binding obligation to enforce foreign liability judgments. The U.S. is not a signatory of any bilateral or multilateral reciprocity agreement that would compel such enforcement.¹⁵⁵ Instead, under general principles of customary international law, foreign courts examine foreign judgments for jurisdiction, notice, fraud, and consistency with the forum's public policy, and judgments involving multiple or punitive damages are generally less likely to be enforced.¹⁵⁶ The public-policy examination is the crux. If the foreign country does not apply strict liability to AI harm or does not allow punitive damages to account for uninsurable risks, its courts are unlikely to cooperate in enforcing tort judgments that depend on those features. The discovery process might also be quite difficult for foreign models, as foreign governments would need to cooperate to enforce subpoenas.

That said, the size of the American market does offer U.S. policymakers substantial leverage. Foreign AI developers are likely to be reluctant to lose access to profit opportunities in selling to customers in the U.S., and U.S. policymakers can certainly attach conditions on market access, including liability insurance or bonding requirements that would ensure that assets under U.S. jurisdiction are available to pay out any claims. This might not solve discovery-related issues, but those could be addressed via shifting the burden of proof when the barrier to proving causation or another key element of a claim is the lack of cooperation from the defendant's host government. Nonetheless, there are important limits to market access as a leverage point. If U.S.-China relations continue to degrade and their economies decouple, then U.S. policymakers will have fewer tools to influence the decisions of Chinese AI developers and deployers. The U.S. also runs a significant trade surplus in digital services, which may also limit its market power in this domain.¹⁵⁷ Importantly, this failure mode applies to any domestic AI regulatory program.

On some level, there is likely no substitute for global coordination to mitigate AI risk. This coordination could take the form of bilateral or multilateral treaties providing for reciprocal

¹⁵⁵ Enforcement of Judgments, <https://travel.state.gov/content/travel/en/legal/travel-legal-considerations/international-judicial-assst/Enforcement-of-Judges.html> (last visited Aug. 4, 2026).

¹⁵⁶ *Id.*

¹⁵⁷ *What Drives the U.S. Services Trade Surplus? Growth in Digitally-Enabled Services Exports*, COUNCIL OF ECONOMIC ADVISERS (June 10, 2024).

enforcement of liability judgments reached by courts in other parties to the treaty. But this approach is only likely to be attractive if countries adopt similarly protective liability regimes. Since the U.S. relies already more heavily on tort liability than other jurisdictions and my proposals would substantially strengthen U.S. tort law in the AI domain, this may be a substantial barrier. An alternative might be for countries to agree to adopt AI risk mitigation measures of comparable effectiveness. In earlier work on climate change mitigation, for instance, I developed a carbon price-equivalent metric that could be used to compare the stringency of qualitatively different climate change mitigation policies across jurisdictions.¹⁵⁸ If a similar metric could be developed for AI risk, countries could commit to enacting policies that score at some specified minimum on this metric. Alternatively, any country or coalition could decide to punish or reward other countries, using the range of carrots and sticks available in diplomacy, for enacting stronger AI risk mitigation policies. One promising variant of this approach is the Chips for Peace framework, under which the U.S. would share access to the benefits of AI with countries that cooperate in enforcing strong safety regulation and preventing the proliferation of AI systems with potentially dangerous capabilities.¹⁵⁹

The potential for misuse of AI systems by government officials is another problem which liability is poorly positioned to address. By default, governments benefit from sovereign immunity. Even in cases where sovereign immunity is waived, liability tends to be a weak tool for influencing the behavior of senior government officials, who respond primarily to political rather than economic incentives. Moreover, highly capable AI systems hold the potential to erode some of the most important safeguards of pluralistic liberal democracy. For instance, a key reason why it is difficult to execute a successful military coup in the U.S. is that it would require many thousands of troops, throughout the chain of command, to give and obey illegal orders. If many military functions could be automated using advanced AI systems, the constitutional order might be much more vulnerable. Similarly, AI could greatly advance the capacity of governments to surveil their populations, making it easier to consolidate power.¹⁶⁰

¹⁵⁸ Gabriel Weil, *The Carbon Price Equivalent: A Metric for Comparing Climate Change Mitigation Efforts Across Jurisdictions*, 125 DICK. L. REV. 475 (2021).

¹⁵⁹ Cullen O’Keefe, *Chips for Peace: How the U.S. and Its Allies Can Lead on Safe and Beneficial AI*, LAWFARE (July 10 2024, 9:38 AM), <https://www.lawfaremedia.org/article/chips-for-peace-how-the-u.s.-and-its-allies-can-lead-on-safe-and-beneficial-ai>.

¹⁶⁰ See generally Martin Beraja, Andrew Kao, David Yang & Noam Yuchtman, *AI-tocracy*, 138 Q. J. ECON. 1349 (2023), <https://academic.oup.com/qje/article/138/3/1349/7076890>.

This is currently another unsolved research and governance problem, but one promising research agenda, law-following AI, examines directly programming AI systems to obey a set of laws.¹⁶¹ While much remains to be fleshed in terms of both legal theory and technical implementation, requirements that powerful AI systems be programmed to follow the law could substantially mitigate the risks laid out in the previous paragraph. To be sure, a determined government with authoritarian inclinations could remove any law-following safeguards or arrange for the training of alternative systems without those constraints. Nonetheless, law-following safeguards could be an important tool for legislatures, courts, and other authorities to restrain the centralizing tendencies of powerful AI systems capable of automating many executive branch functions. Law-following AI would be of particular importance in scenarios where major AI labs are nationalized, as some observers expect as systems get more powerful.¹⁶² That said, the law-following AI research agenda is embryonic and may not live up to its promise, so other avenues for solving these governance challenges are also worth exploring.

Finally, let's turn to externalities that are unlikely to give rise to enforceable liability judgments. These include uninsurable risks for which warning shots are unlikely, legally non-compensable harm like political misinformation and election interference, and actions whose risks may only be realized via highly attenuated causal chains. There is also an adjacent category of harm for which enforceable liability judgments are plausible, but are probably not the best risk mitigation tool, because the arguments laid out in part II do not fully apply. The primary example of this is cybersecurity breaches antecedent to misuse. For reasons explained earlier in this section, the incentives provided by liability may be inadequate to drive strong cybersecurity for frontier AI systems. Since inadequate cybersecurity *also* contributes to downstream risks that are unlikely to trigger liability, this fact does not have strong implications for AI governance. The same measures, direct regulation and government provision of cybersecurity services, that may

¹⁶¹ Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay & Christoph Winter, *Law-Following AI: Designing AI Agents to Obey Human Laws*, 94 FORDHAM L. REV. 57 (2025); LawAI Workstreams & Research Directions (shared) 6, GOOGLE DOCS, <https://docs.google.com/document/d/1JldBwEDhKgNH-DPZfv7O0bVpS1dNdaz8-OikZY01nSY/edit#heading=h.q5ac51ugtu1c> (last visited Aug. 4, 2026); see generally Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 VAND. J. ENT. & TECH. L. 133 (2020), <https://scholarship.law.vanderbilt.edu/jetlaw/vol19/iss1/5/>; see generally Mark Chinen, *Law and Autonomous Machines: The Co-evolution of Legal Responsibility and Technology* (2020).

¹⁶² E.g., Deric Cheng & Corin Katzke, *Soft Nationalization: How the US Government Will Control AI Labs*, AGI, 2024, at 2, <https://agi-rsa.com/index.php/agi/article/view/10931/7759>.

be warranted by geopolitical considerations or concerns about race dynamics would also address more direct harm pathways.

The most significant potential shortcoming of a liability-based AI governance regime is the class of uninsurable risks for which near-miss punitive damages offer an insufficient deterrent, either because warning shots are too unlikely or because the risk is so catastrophic and likely, say, a 1% risk of human extinction, that no feasible damages award could compel AI developers and deployers to internalize it. In this sense, the approval regulation backstop described in the previous section may be best thought of as a complementary policy, rather than a supportive policy. It was outlined in the previous section only because its implementation would be tightly linked with the liability insurance-based licensing regime that is best understood as a supportive policy.

Factual and proximate cause are essential elements of most tort liability claims, including all negligence and strict liability claims. To establish factual cause, plaintiffs generally¹⁶³ need to prove by a preponderance of the evidence that the defendant's breach of duty, the defect in the defendant's product, or the defendant's engagement in an abnormally dangerous activity was the but-for cause of the plaintiff's injury. That is, had the defendant exercised reasonable care, the product not been defective, or the defendant not engaged in the abnormally dangerous activity, the plaintiff's injury would have more likely than not been prevented.

Proximate cause doctrine is less straightforward, but the general idea is that defendants should only be held liable for the foreseeable consequences of their actions. Imagine driver α negligently cuts off driver β in traffic, causing driver β to slam on her brakes to avoid a collision. Driver β successfully avoids the collision, but is delayed by five seconds. A mile down the road, driver β is sideswiped by driver γ in an intersection that they would have cleared safely but for the five-second delay. Driver α breached his duty of care, and that breach was a but-for cause of driver β 's injury, but no court would hold driver α liable for those injuries, and not just because driver β or driver γ may have been at fault for the later collision. The collisions a mile down the road is simply not the sort of risk that made it unreasonable for driver α to cut off driver β . Indeed, driver α 's actions could just as easily have prevented driver β 's later collision as caused

¹⁶³ Ernest Weinrib, *A Step Forward in Factual Causation*, 38 MOD. L. REV. 518, 518-520 (1975), <https://onlinelibrary.wiley.com/doi/pdfdirect/10.1111/j.1468-2230.1975.tb01428.x>.

it. The upshot of these two requirements is that courts are often reluctant to hold defendants liable for harm involving complex, murky, and attenuated causal chains.

Two potential scenarios have come up in discussions of this issue. One scenario involves a malicious foreign actor fine-tuning an AI system before using it to commit harm. This could either be an open-weights model or a proprietary system subject to a cybersecurity breach that leaks the model weights. After fine-tuning, it may be difficult to determine which base model was used to harm the plaintiff. This problem could be mitigated by traceability proposals like the watermarking requirements discussed in section A on supportive policies. But an actor with access to the model weights could likely train away any watermarks if they were motivated to do so. Proprietary models could be required to disclose cybersecurity, with stiff penalties if they are caught hiding them, but it is possible that a cybersecurity breach could occur without the knowledge of the model developer. To address this, cybersecurity breach disclosure requirements might be paired with a requirement to implement cybersecurity breach trip wires. Along these lines, Cullen O’Keefe has proposed that advanced AI system developers be required to create a “honeypot” by placing a large amount of cryptocurrency in a wallet that regulators can monitor.¹⁶⁴ The existence of these accounts would be announced to the world so that hackers that are able to access the model weights would have an incentive to reveal the security breach by removing the crypto.¹⁶⁵

Open-weights models still pose a substantial traceability problem, however, since no cybersecurity breach is needed to train away any safeguards against misuse or watermarks. To address this problem, the backstop approval regulator might be empowered to disallow open-weights model deployment for sufficiently powerful systems. If the system has demonstrated the ability to cause large and likely untraceable harm, then the regulator would have a strong basis for concluding that even a robust strict liability regime with punitive damages and mandatory liability insurance would be inadequate to ensure adequate precautions. Alternatively, permission to release the weights of a powerful model could be conditioned on the developer making an upfront contribution to a misuse victim-compensation fund, which would pay out claims in

¹⁶⁴ Cullen O’Keefe, *Crypto Honeypots for Evidence of Model Weight Theft*, JURAL NETWORKS (Jan. 9, 2025), https://juralnetworks.substack.com/p/crypto-honeypots-for-evidence-of?utm_source=post-email-title&publication_id=2133027&post_id=147758138&utm_campaign=email-post-title&isFreemail=true&r=u45z&triedRedirect=true&utm_medium=email.

¹⁶⁵ *Id.*

misuse cases where it is unclear which base model was used. Ideally, this would function as a sort of ex ante AI risk tax, with the payment commensurate with the assessed risk generated by the system. This would pose greater informational challenges than ex post liability, but may be the best option for governing potentially untraceable open-source models.

The other scenario involves interactions between multiple AI systems. Ashley Deeks analyzes this issue in the military context, where the interaction is between two automated fleets of armed autonomous underwater vehicles (AUVs). In Deeks's scenario:

State B's AUVs have been trained to conduct acoustic jamming against any adversarial active sonar it encounters. In turn, State A's AUVs have been trained to treat as hostile any acoustic jamming they encounter and to fire warning shots at the source of that jamming. State B's AUVs have been built to defend themselves against any object that shoots in its direction, and thus fire missiles at each of State A's AUVs, sinking them.

While this scenario would be unlikely to give rise to tort liability even if we could identify the responsible party, similar scenarios arise involving non-military automated systems or harm to third parties. But tort law does have tools for handling such cases. We can tweak Deeks' scenario to make this clear. Imagine a private submersible, owned and operated by Citizen C, is caught in the crossfire between State A and State B's AUV fleets. Standard factual cause analysis would conclude that the deployment of *both* State A and State B's AUV fleets was necessary for Citizen C's injury, so both of their decisions to deploy were but-for causes. We would still need to apply the relevant duty/breach or strict liability analysis to all three actors, but the involvement of multiple systems does not defeat liability.

If we imagine a structurally similar situation involving three private parties, then Citizen C's recovery would depend on the applicable liability regime and an assessment of his own fault.¹⁶⁶ The two main liability regimes are joint and several liability and several liability with apportionment. Under both regimes, the fact finder would be tasked with determining a fault allocation between the defendant and, if applicable, the plaintiff.¹⁶⁷ Assuming C behaved reasonably and that A and B were in breach of their duty of care or subject to strict liability, then joint and several liability would enable C to recover in full from either A or B, who would then

¹⁶⁶ Kathleen O'Connor & Gregory Sreenan, *Apportionment of Damages: Evolution of a Fault-Based System of Liability for Negligence*, 61 J. AIR L. & COM. 365, 370-3 (1995), <https://scholar.smu.edu/cgi/viewcontent.cgi?article=1396&context=jalc>.

¹⁶⁷ O'Connor & Sreenan *supra* note 166, at 373-74.

have the recourse of suing their joint tortfeasor for contribution of their fault-apportioned share.¹⁶⁸ In a jurisdiction that follows several liability with apportionment, the plaintiff would only be allowed to recover each defendant's fault apportioned share.¹⁶⁹ In any case, juries (and judges in bench trials) are typically deemed competent to allocate fault in such cases.¹⁷⁰ More complex cases might obscure factual cause, but that is also a problem that has long been faced by the tort system. It is not clear that the AI context presents any novel challenges in terms of assigning responsibility in cases involving multiple potential causes.

Finally, there are classes of AI risk that are unlikely to give rise to a tort liability claim, due to the nature of the harm. Prominent examples include algorithmic generated misinformation and election interference, but there are likely other examples. Scholars,¹⁷¹ media reports,¹⁷² and senior government officials¹⁷³ are already warning that deepfakes and other AI-generated misinformation are undermining public health¹⁷⁴ and election integrity.¹⁷⁵ Some scholars also suggest AI can provide part of the solution, by automating fact checking and content moderation.¹⁷⁶ Other scholars argue that the real problem is too much demand for misinformation, not too much supply.¹⁷⁷ That is, there are many people that want to believe false

¹⁶⁸ Robert Franks, *Contribution and Indemnity among Tortfeasors*, 31 MONT. L. REV. 69, 78-9 (1969), <https://scholarworks.umt.edu/cgi/viewcontent.cgi?article=1794&context=mlr>.

¹⁶⁹ Jonathan Hoffman, *Claim Splitting in the New World of Several Liability and Personal Jurisdiction*, 86 J. AIR L. & COM. 377, 388 (2021), <https://scholar.smu.edu/cgi/viewcontent.cgi?article=4181&context=jalc>.

¹⁷⁰ Daniel Levi, *A Comparison of Comparative Negligence Statutes: Jury Allocation of Fault—Do Defendants Risk Paying for the Fault of Nonparty Tortfeasors?*, 76 WASH. U. L.Q. 407, 408 (1998), <https://journals.library.wustl.edu/lawreview/article/6943/galley/23776/view/>.

¹⁷¹ Angela Yang, *AI image misinformation has surged, Google researchers find*, NBC NEWS (May 29, 2024), <https://www.nbcnews.com/tech/tech-news/ai-image-misinformation-surged-google-research-finds-rcna154333>.

¹⁷² Pranshu Verma, *The rise of AI fake news is creating a 'misinformation superspreader'*, WASH. POST (Dec. 17, 2023), <https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/>.

¹⁷³ Ali Swenson & Kelvin Chan, *AI-created election disinformation is deceiving the world*, AP NEWS (Mar. 14, 2024), <https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd>.

¹⁷⁴ *How Can We Tackle AI-Fueled Misinformation and Disinformation in Public Health?*, Center on Emerging Infectious Diseases, <https://www.bu.edu/ceid/2024/04/25/how-can-we-tackle-ai-fueled-misinformation-and-disinformation-in-public-health/> (last visited Aug. 4, 2026).

¹⁷⁵ Ali Swenson, *Tests find AI tools readily create election lies from the voices of well-known political leaders*, AP NEWS (May 31, 2024), <https://apnews.com/article/artificial-intelligence-audio-voice-cloning-elections-2024-2500813b642169478c27c168aab1b3e3>.

¹⁷⁶ Jieun Shin, *AI and Misinformation*, 2024 Dean's Report (2024), <https://2024.jou.ufl.edu/page/ai-and-misinformation> (last visited Aug. 4, 2026).

¹⁷⁷ Tyler Cowen, *Too Much Misinformation? The Issue Is Demand, Not Supply*, BLOOMBERG (Oct. 3, 2023, 7:30 AM), <https://www.bloomberg.com/opinion/articles/2023-10-03/campaign-2024-will-ai-generated-misinformation-be-a-big-problem>; Ilya Somin, *The Demand for Political Misinformation is a Bigger Problem than the Supply - Even in the Age of AI*, REASON.COM (Oct. 5, 2023, 10:30 AM), <https://reason.com/volokh/2023/10/05/the-demand-for-political-misinformation-is-a-bigger-problem-than-the-supply-even-in-the-age-of-ai/>; David Wemer, *How do you*

narratives and will latch on to any information that supports those narratives, including low-tech versions of misinformation.¹⁷⁸ This phenomenon is particularly pronounced on political issues, since individual voters have an extremely low chance of casting a decisive vote, and thus little incentive to seek out objective information sources and form accurate beliefs.¹⁷⁹ On this view, AI's augmentation of the quality and volume of misinformation is unlikely to substantially worsen the problem.¹⁸⁰ I am somewhat more persuaded by the latter perspective. Nonetheless, it is important to acknowledge that, to the extent AI misinformation warrants a policy solution, liability is unlikely to provide it, particularly with regard to political misinformation.

One promising approach to combating misinformation is truth bounties, a contractual mechanism under which speakers pledge money to back the credibility of their communications and any person can claim the pledged funds by proving the communication's falsity to a private arbitrator.¹⁸¹ Consumers of information would know which communications carry a bounty and whether they had been challenged, giving those who value accuracy a guide to the information ecosystem and good-faith communicators a means of credibly signaling their epistemic virtue. This would do little to address the demand side of misinformation, but that is not really an AI governance problem, and thus beyond the scope of this Article. Other ideas discussed above, like watermarking of AI content, may also be useful in this context. Watermarking would be especially useful if it can be tied to particular sources of information with reputations for accuracy.

Finally, consider election interference. AI systems could be used to hack voting machines, vote-transmission systems, and voter-registration databases, potentially creating chaos and casting the results of an election into doubt. This sort of scenario would not typically give rise to tort liability, at least for the creators of the AI system. The end user would be subject to criminal liability, but there is no clear pathway to accountability for the model developer or intermediate deployer. Nonetheless, this sort of risk is readily amenable to a policy response in the spirit of ex post liability. That is, model developers and intermediate deployers whose actions

defeat disinformation? Tackle demand, not just supply, ATLANTIC COUNCIL (Mar. 5, 2019), <https://www.atlanticcouncil.org/blogs/new-atlanticist/how-do-you-defeat-disinformation-tackle-demand-not-just-supply/>.

¹⁷⁸ Cowen, *supra* note 177.

¹⁷⁹ Somin, *supra* note 177.

¹⁸⁰ Cowen, *supra* note 177.

¹⁸¹ See generally Yonathan A. Arbel & Michael D. Gilbert, *Truth Bounties: A Market Solution to Fake News*, 102 N.C. L. REV. 509 (2024), <https://scholarship.law.unc.edu/cgi/viewcontent.cgi?article=6960&context=nclr>.

enable such scenarios could be subject to civil penalties or fines that reflect the social costs of their actions. Liability insurance mandates could include a requirement that such non-tort liability be covered by qualifying policies. It will also likely make sense to harden election integrity as a target through enhanced cybersecurity and maintenance of paper trails, and otherwise tweak and bolster the existing laws and institutions designed to protect election integrity from conventional threats.

C. Substitutionary Policies

My claims in part II about the relative benefits of ex post liability were largely predicated on a liability system that is much stronger than the status quo. In particular, the status quo liability regime will likely apply a negligence standard rather than strict liability to most cases of AI harm, will require a showing of defendant malice or recklessness to qualify for punitive damages, and lacks liability insurance requirements. This means that AI developers and deployers are likely to externalize much of the risk they create. In many cases, this will be because the plaintiff is unable to point to some precautionary measure that a reasonable AI developer or deployer would have taken to prevent the plaintiff's injury. As I've argued in prior work, merely adopting all the affirmative precautionary measures that are standard in the AI industry is unlikely to be sufficient to protect society from the risks of AI. And many of the measures that are most likely to mitigate AI risk, like investing in developing new alignment and safety techniques and delaying deployment of potentially dangerous systems, are unlikely to be required by standard negligence analysis. This is either because they entail pushing forward the safety frontier and thus exceeding the requirement of the reasonable person standard or because they fall outside of the narrow scope of the standard negligence inquiry, which fails to regulate the volume or mode of risky activities.¹⁸²

In a smaller, but potentially much more consequential, number of cases, AI developers and deployers may escape liability because they are judgment-proof. Defendants might be judgment-proof because they are corporations whose investors benefit from limited liability and the size of the damage awards exceeds the value of the company and any liability insurance coverage. Or the harm might be so large as to exceed any plausible ability to pay, even with the

¹⁸² Gabriel Weil, *Efficiency, Fairness, and the Externalization of Reasonable Risks: The Problem With the Learned Hand Formula*, 75 S.C. L. REV. 155 (2023), <https://sclawreview.org/article/efficiency-fairness-and-the-externalization-of-reasonable-risks-the-problem-with-the-learned-hand-formula/>.

maximum feasible insurance policy. In even more severe cases, the legal system may no longer be functional or no humans may be alive to sue or be sued. Ideally, the first sort of scenario would be addressed by liability insurance requirements that scale with model capabilities, as described in section A of this part. The other scenarios, those involving uninsurable catastrophes, cannot be addressed via ex post compensatory damages. But that does not mean liability is impotent. Instead, AI developers and deployers can be given incentives to mitigate uninsurable risks by allowing punitive damages, with no requirement that the plaintiff prove a defendant's malice or recklessness, in cases of practically compensable harm that are associated with uninsurable risk.

That said, as discussed in the previous section, this mechanism fails where uninsurable risks are too large or warning shots too unlikely. But if the requirement for defendant malice or recklessness is not relaxed in cases of AI harm that serve as warning shots for uninsurable risks, then tort law would fail even for lower-probability, uninsurable risks with abundant warning shots. To be sure, there might be some cases where plaintiffs could meet the high burden of proving that the defendant acted with malice or recklessness, a much higher bar than proving ordinary negligence. But the prospect of punitive damages in such cases would only provide AI developers and deployers with incentives to refrain from the most egregious conduct; it would stop far short of aligning their incentives with social welfare.

Finally, if liability insurance is not required, the class of risks effectively covered by liability would shrink still smaller. AI developers and deployers have no pecuniary reason to account for incremental risks of harm larger than the value of their total enterprise, including any voluntary insurance coverage. While frontier AI companies are valued at tens of billions of dollars,¹⁸³ moreover, a significant portion of these valuations is premised on the possibility that frontier AI systems will safely generate much greater value in the future.¹⁸⁴ This implies that the liquidation value of these companies is likely far smaller. Leaders at AI companies may perceive greater risk to the value of their equity from moving too slowly compared to risk of going under after being held liable for catastrophic harm.

¹⁸³ Cade Metz & Tripp Mickle, *OpenAI Completes Deal That Values the Company at \$80 Billion*, N.Y. TIMES (Feb. 16, 2024), <https://www.nytimes.com/2024/02/16/technology/openai-artificial-intelligence-deal-valuation.html>.

¹⁸⁴ Gergely Orosz, *What is OpenAI, Really?*, THE PRAGMATIC ENGINEER (Nov. 23, 2023), <https://newsletter.pragmaticengineer.com/p/what-is-openai>.

The analysis above suggests that the default liability regime is deeply inadequate. If the legal and political barriers to enacting a robust liability regime prove insurmountable, the arguments in Part II regarding the superiority of ex post liability as a governance tool largely fall by the wayside. In that case, we cannot avoid confronting the epistemic and organizational challenges of ex ante governance tools. The remainder of this section sketches out what other policy instruments come back into play in scenarios where the liability regime falls short in at least one respect.

First, consider a scenario in which most states do adopt strict liability, either via legislation or judicial application of the abnormally dangerous activities doctrine,¹⁸⁵ and require liability insurance that scales with potentially dangerous model capabilities, but do not expand the availability of punitive damages in order to mitigate uninsurable risks. Under such a regime, ex post liability would handle insurable risks quite well, but totally break down for uninsurable risks. In section B of this part, I outlined a backstop approval regime that would apply in scenarios where the regulator can prove that the system presents risks that cannot be addressed by ex post liability, even with insurance requirements and punitive damages in near miss cases. In the absence of punitive damages, it would be sensible to expand this authority, allowing regulators to disapprove models or model deployment plans if they can prove the model generates uninsurable risks that exceed the expected social benefits flowing from the system's deployment. It would probably make sense to pair this approval regime with a set of prescriptive rules targeted at mitigating uninsurable risks. Such rules might include requirements for safety testing,¹⁸⁶ red teaming,¹⁸⁷ and third-party audits,¹⁸⁸ tighter restrictions on open source release of

¹⁸⁵ There are other doctrinal pathways to strict liability, including vicarious liability and products liability. I address these doctrinal issues in detail in a forthcoming paper, tentatively titled *Abnormally Dangerous Algorithms*. They are also discussed in Weil, *supra* note 10. My view is that products liability is inadequate, extension of the abnormally dangerous activities doctrine to frontier AI development is essential, and vicarious liability should play an important supporting role in cases involving broadly deployed AI agents.

¹⁸⁶ Cullen O'Keefe, Jade Leung & Markus Anderljung, *How technical safety standards could promote TAI safety*, EFFECTIVE ALTRUISM FORUM (Aug. 8, 2022), <https://forum.effectivealtruism.org/posts/zvbGXCxc5jBowCuNX/how-technical-safety-standards-could-promote-tai-safety>.

¹⁸⁷ Jakob Mökander et al., *Auditing large language models: a three-layered approach*, 4 AI AND ETHICS 1085, 1087 (2023), <https://link.springer.com/article/10.1007/s43681-023-00289-2>.

¹⁸⁸ Abeba Birhane et al., *AI auditing: The Broken Bus on the Road to AI Accountability*, arXiv.org (Jan. 5, 2024), <https://arxiv.org/abs/2401.14462>.

advanced AI systems,¹⁸⁹ and requirements that AI developers follow particular model development procedures or include specified design features in their systems.¹⁹⁰ In selecting among these interventions, it would make sense to prioritize those with the largest impact on uninsurable risks, since insurable risks would still be well-governed by the liability regime.

Next, consider a scenario with strict liability, but no liability insurance requirements or expansion of punitive damages to address uninsurable risks. Under this regime, AI developers and deployers would be judgment-proof for many harms arising from insurable risks, not just uninsurable risks,¹⁹¹ though their incentives would remain well-aligned for harms too small to push the responsible party into bankruptcy. In such a scenario, the approval regulator should consider the capitalization of the developer and potential deployers in deciding whether to grant a deployment license, and applicants could strengthen their applications by voluntarily putting up a bond or taking out an insurance policy. The scope of both approval authority and prescriptive regulation would need to be broader, extending even to large insurable risks.¹⁹² Governments would also need to invest more in the in-house expertise required to evaluate model safety, since regulators would be assessing the safety cases that private insurers would otherwise vet. The political will for regulation of this strength may be lacking until existing AI systems clearly demonstrate dangerous capabilities. If it does arrive, it will be essential to move quickly to build out these regimes, and policymakers could lay some of that groundwork now, in the spirit of what Christoph Winter and Charlie Bullock call radical optionality, by establishing the information channels and legal authorities that rapid action would require.¹⁹³

¹⁸⁹ Elizabeth Seger et al., Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives, arXiv.org (Sept. 29, 2023), <https://arxiv.org/abs/2311.09227>.

¹⁹⁰ Ryan Greenblatt et al., AI Control: Improving Safety Despite Intentional Subversion, arXiv.org (Dec. 12, 2023), <https://arxiv.org/abs/2312.06942>.

¹⁹¹ Lynn M. LoPucki, *The Death of Liability*, 106 YALE L.J. 1 (1996); Brian Galle, *In Praise of Ex Ante Regulation*, 68 VAND. L. REV. 1715, 1724-5 (2015), <https://scholarship.law.vanderbilt.edu/cgi/viewcontent.cgi?article=1207&context=vlr>.

¹⁹² Prescriptive regulations set specific rules that must be followed in conducting an activity, such as detailed requirements for workplace safety equipment or emissions control technologies. These regulations dictate exactly how something must be done. In contrast, approval regulation requires obtaining permission before engaging in an activity, like securing a license to practice medicine or getting pre-approval for a new pharmaceutical drug. While prescriptive regulations focus on how an activity is conducted, approval regulations control whether an activity can be undertaken at all, creating a gatekeeping function that occurs before the activity begins.

¹⁹³ Christoph Winter & Charlie Bullock, *Radical Optionality: Governing Transformative AI Under Uncertainty*, INST. FOR L. & AI (Apr. 23, 2026), <https://radical-optionality.ai/>.

Now, consider the reverse scenario where legislatures do act to impose liability insurance requirements, but both courts and legislatures decline to shift from a negligence regime to strict liability or expand availability of punitive damages. Under this regime, all insurable risks would be at least partly addressed, but AI developers and deployers would lack incentives to reduce risks in ways that fall outside the scope of the negligence inquiry or go beyond the requirements of reasonable care. Recall that the scope of the breach analysis in negligence cases is quite narrow. If a driver runs over a pedestrian with their car, courts do not ask whether the expected net benefits of the driving trip to the driver and any passengers exceed the injury risks that the trip generated for third parties. That is, negligence analysis fails to regulate the volume of risky activity.¹⁹⁴ Similarly, if a driver strikes a pedestrian with an SUV, inflicting greater injuries than they would have with a compact sedan or a bicycle, courts do not ask whether the net benefits to driver and any passengers of using the SUV justify the greater external risk. That is, negligence also fails to regulate the mode of risky activity.¹⁹⁵ Even within the scope of the negligence inquiry, merely adopting the established best safety practices that the reasonable-care standard will likely be understood to require may offer inadequate protection. In this scenario, non-liability governance tools would be responsible both for mitigating uninsurable risks and for inducing risk mitigation measures that go beyond or fall outside the scope of the negligence inquiry. The backstop approval authority and prescriptive rules described above would again be warranted, with the prescriptive rules now also aimed at reducing insurable risks via means the reasonable-care standard does not require. These requirements could be regularly updated as the leading labs develop new safety practices. Instead of waiting for a practice to achieve widespread enough adoption to be incorporated into the reasonable-care standard, legislators and regulators could proactively require its universal adoption. Regulations could also require that frontier AI developers dedicate a specified share of staff time, computing, and other resources to alignment and safety research. A resource mandate of this kind would not perfectly align developers' incentives, but it would empower the safety-conscious actors within AI companies, and the profits from frontier AI systems could be taxed to fund public alignment and safety research to inform both prescriptive and approval regulation.

¹⁹⁴ Steven Shavell, *Strict Liability Versus Negligence*, 9 J. LEGAL STUD. 1, 2 (1980), <https://www.journals.uchicago.edu/doi/10.1086/467626>.

¹⁹⁵ Weil, *supra* note 182.

Finally, consider the default liability regime, with only negligence and, in some cases, products liability available, no liability insurance requirements, and highly constrained availability of punitive damages. In this scenario, the liability regime would address only the subset of insurable risks for which compensatory damages awards would be practically enforceable without insurance requirements, and only insofar as the reasonable-care standard, or the functionally similar reasonable alternative-design standard for product design defects, mitigates them. This leaves the broadest scope for the tools canvassed above. An approval regulator could be empowered to bar training and deployment of models that generate risks larger than the responsible party's demonstrated ability to pay and that exceed the expected external social benefits of the system, and the prescriptive rules, resource mandates, and safety-research taxes described above would be warranted whenever they offer cost-effective mitigation of risks that exceed the responsible parties' practical ability to pay, fall outside the scope of the reasonable-care standard, or would not be required under it.

IV. Considerations on Preemption and Displacement

It is possible that any major federal authorization of new AI regulations will also include a provision expressly preempting state tort law.¹⁹⁶ Indeed, several iterations of the One Big Beautiful Bill Act, passed via budget reconciliation in July of 2025, included a provision that would have barred states from enforcing certain laws or regulation concerning AI systems.¹⁹⁷ The provision underwent a series of changes and was ultimately stripped out of the bill before it passed the Senate.¹⁹⁸ The original House-passed version was quite broad and would have directly preempted state AI-related laws for 10 years.¹⁹⁹ In the Senate, the provision was modified to serve as a condition on new Broadband Equity, Access, and Deployment (BEAD) grants to

¹⁹⁶ Dean W. Ball & Alan Z. Rozenshtein, *Congress Should Preempt State AI Safety Legislation*, LAWFARE (June 17, 2024, 2:00 PM), <https://www.lawfaremedia.org/article/congress-should-preempt-state-ai-safety-legislation>; Jim Harper, *AI Regulation, Preemption, and Democracy*, AMERICAN ENTERPRISE INSTITUTE (June 23, 2024), <https://www.aei.org/technology-and-innovation/ai-regulation-preemption-and-democracy/>; Brendan Bordelon, *Obernolte Urges Gov. Newsom to Veto California's AI Safety Bill*, POLITICO (Sept. 17, 2024, 2:38 PM), <https://www.politico.com/live-updates/2024/09/17/politico-ai-and-tech-summit/obernolte-veto-california-ai-bill-00179622>.

¹⁹⁷ Mackenzie Arnold & Charlie Bullock, *The AI Moratorium—the Blackburn Amendment and New Requirements for “Generally Applicable” Laws*, INST. FOR L. & AI (2025), <https://law-ai.org/the-ai-moratorium-the-blackburn-amendment-and-new-requirements-for-generally-applicable-laws/> (last visited Aug. 4, 2026).

¹⁹⁸ *Id.*

¹⁹⁹ Gabriel Weil, *The House Reconciliation Bill's AI Preemption Clearly Violates the Byrd Rule*, LAWFARE (May 30, 2025), <https://www.lawfaremedia.org/article/the-house-reconciliation-bill-s-ai-preemption-clearly-violates-the-byrd-rule>.

states, in order to comply with the Byrd Rule, which requires that all provisions in budget reconciliation bills have an impact of federal spending or revenues that is not merely incidental. The provision was also later modified to clarify that it did not bar states from enforcing generally applicable laws.²⁰⁰ It is unclear how this would have applied to judicial application of the abnormally dangerous activities doctrine to frontier AI development, but it very likely still would have barred states that accepted newly appropriated BEAD funds from enforcing any special AI liability legislation that created strict liability, liability insurance requirements, or punitive damages. While this particular effort fell short, there is still a meaningful constituency looking for ways to block at least some forms of state AI legislation, potentially including liability rules.²⁰¹

That constituency now has executive-branch support. Executive Order 14,365, signed December 11, 2025, directs the Attorney General to establish a litigation task force charged with challenging state AI laws, instructs the Commerce Department to identify state AI laws it deems onerous, and conditions Broadband Equity, Access, and Deployment funding on state non-enforcement, with carve-outs for areas such as child safety and state procurement.²⁰² The White House followed in March 2026 with its National Policy Framework for Artificial Intelligence, a set of legislative recommendations urging Congress to preempt state AI laws the administration characterizes as unduly burdensome while preserving generally applicable law.²⁰³ Draft legislation has begun to give that recommendation content. Senator Blackburn's TRUMP AMERICA AI Act discussion draft would impose a duty of care on AI developers enforced by the Federal Trade Commission, create federal claims for defective design and failure to warn, and preempt state regulation of frontier developers' catastrophic-risk management, while expressly disclaiming preemption of generally applicable law, including the common law.²⁰⁴ The

²⁰⁰ Arnold & Bullock, *supra* note 197.

²⁰¹ Hayden Field, Trump Is Bringing Back the AI Law Moratorium, THE VERGE, July 23, 2025, <https://www.theverge.com/ai-artificial-intelligence/712537/trump-ai-action-plan-white-house-ai-law-moratorium> (last visited Aug. 5, 2026).

²⁰² Exec. Order No. 14,365, 90 Fed. Reg. 58,499 (Dec. 11, 2025) (Ensuring a National Policy Framework for Artificial Intelligence).

²⁰³ WHITE HOUSE, *National Policy Framework for Artificial Intelligence* (Mar. 20, 2026), <https://www.whitehouse.gov/releases/2026/03/president-donald-j-trump-unveils-national-ai-legislative-framework/>.

²⁰⁴ TRUMP AMERICA AI Act (Sen. Blackburn, discussion draft Mar. 18, 2026) (not introduced as of Aug. 4, 2026); see Press Release, Sen. Marsha Blackburn, Blackburn Releases Discussion Draft of National Policy Framework for Artificial Intelligence (Mar. 18, 2026), <https://www.blackburn.senate.gov/2026/3/technology/blackburn-releases-discussion-draft-of-national-policy-framework-for-artificial-intelligence/3b3b6458-b6c7-478b-9859-374949586765>.

AI LEAD Act, introduced by Senators Durbin and Hawley, would instead add a federal products-liability framework for AI systems, with claims for defective design and failure to warn and a bar on terms-of-service liability waivers.²⁰⁵ The Great American Artificial Intelligence Act discussion draft, released in June 2026, went further, pairing comprehensive federal development-conduct standards for frontier models with preemption of state laws specifically regulating model development while saving only remedies “existing under the common law,” a formulation that arguably freezes preserved claims as of enactment.²⁰⁶ The FRONTIER Act, its bipartisan successor introduced in July 2026, improves the design. It imposes tiered transparency, audit, and incident-reporting requirements, saves laws of general applicability, and confines preemption to defined subject areas, though it is open to an interpretation under which there would be incidental preemption of some tort claims. Disappointingly, it also sets frontier developers up as the audit customers, so verifiers win business by accommodating them, the issuer-pays conflict that discredited the credit-rating agencies.²⁰⁷ The treatment of liability is the detail to watch: the leading legislative vehicles so far either preserve common-law claims or add federal ones, and whether those provisions survive the legislative process will determine whether federal preemption arrives coupled to accountability or stripped of it.

Perhaps unsurprisingly, I think this would be unfortunate, unless the bill creates a strong federal liability regime. Nonetheless, it is worth considering the conditions under which this would be a trade worth making. After all, even if, as I argue above, a first-best AI risk mitigation policy regime would have liability at its center, such a regime would require much stronger liability rules than what is likely to emerge by default.²⁰⁸ If faced with a bill that contains regulatory provisions that hold the promise of substantially mitigating AI risk, many in the AI safety community would likely be willing to accept preemption of state tort law, along with preemption of other relevant state policies, as part of that deal. Moreover, there may not be much time to deliberate about whether such a bill would be worth supporting, once it emerges. In this part, I survey the relevant considerations and offer a framework for evaluating such choices.

²⁰⁵ AI LEAD Act, S. 2937, 119th Cong. (2025) (referred to the S. Comm. on the Judiciary).

²⁰⁶ Great American Artificial Intelligence Act of 2026, § 121(c)(1) (discussion draft, June 2, 2026); see Charlie Bullock, *Congress Should Do Something: The Case for (Fixing) the Great American AI Act*, LAWFARE (July 8, 2026), [https://www.lawfaremedia.org/article/congress-should-do-something--the-case-for-\(fixing\)-the-great-american-ai-act](https://www.lawfaremedia.org/article/congress-should-do-something--the-case-for-(fixing)-the-great-american-ai-act).

²⁰⁷ FRONTIER Act, H.R. 9925, 119th Cong. §§ 5, 9 (2026); see Weil, *Don't Let AI Developers Hire Their Own Referees*, *supra* note 108.

²⁰⁸ Weil, *supra* note 10, at 21-44.

Federal preemption of state tort law has two primary effects. First, it eliminates the default liability regime: negligence and products liability, with limited availability of punitive damages and no liability insurance requirements. Second, it cuts off the potential development of state liability regimes to include one or more of the components of the ideal liability regime outlined above: strict liability, punitive damages tied to uninsurable risks, and liability insurance requirements. The foregoing analysis in the previous section highlighted the limitations of the default liability regime, but it does offer some risk mitigation benefits. The reasonable-care standard can function as a catch-all for unreasonably risky actions that are not expressly prohibited by prescriptive regulations, so long as the potential mitigation measures fall within the scope of the negligence inquiry. This is likely to be particularly valuable if the costs of these measures are relatively low (so that they can be justified on the basis of avoiding potential liability for practically compensable harm) and the same measures that mitigate practically compensable harm also tend to mitigate uninsurable risks. The value of a catch-all negligence standard is likely to be highest if the federal regulatory regime lacks a comparably broad reasonable-care provision. By contrast, if the most important mitigation measures for limiting uninsurable risks can only be justified on the basis of those risks or would not be required by the reasonable-care standard, then a negligence regime with limited availability of punitive damages will offer much less risk mitigation. Similarly, the more comprehensive the federal regulatory scheme, the less incremental value a catch-all reasonable-care requirement can offer. While there is substantial uncertainty about these variables, the expected risk mitigation afforded by the default liability regime is sufficiently modest that it is not hard to imagine a regulatory scheme that would offer enough risk mitigation, at low enough cost, to be worth sacrificing it.

The potential future development of state tort regimes is tougher to value. As I argue in prior work, there is a clear doctrinal rationale, in the form of the abnormally dangerous activities doctrine, for adopting a strict liability standard for harm arising from the inherently dangerous nature of training and deploying powerful and poorly understood AI systems. While I do not view this as the likely default evolution of tort doctrine, it is plausible that the courts in one or more states would be persuaded to adopt this relatively minor doctrinal innovation. This one change would fully align the incentives of AI developers and deployers with respect to risks of practically compensable harm. To the extent that the measures that offer the greatest potential mitigation of catastrophic risk would also tend to mitigate risks of practically compensable harm

and are cheap enough to be justified on that basis, this regime offers substantial protection against the full range of AI risks. Since this is probably the most likely scenario, aside from the default condition, further work from technical AI risk experts to characterize the shape of the risk curve would be useful. Similarly, estimates of the likely efficacy of mitigation measures motivated by risks of practically compensable harm in mitigating larger risks would be important to a quantitative assessment of the value of such a regime. While both the likelihood²⁰⁹ of this scenario and its value are difficult to quantify, the prospect of a strict liability regime emerging does raise the bar for AI regulatory legislation for which it would be worth accepting liability preemption.

The other liability reform measure that seems plausible without a major change in the facts on the ground is liability insurance requirements. While this step would require new legislation, it is hard to argue that demonstrating the ability to pay for the harm caused by your systems is an unreasonable burden to put on AI companies. Every U.S. state already requires liability insurance for drivers on public roads, so there is a clear precedent.²¹⁰ In my conversations with state legislators and congressional staff, there is also substantial interest in this idea and a perception that it is within the Overton Window, perhaps more so than strict liability.²¹¹ While liability insurance requirements would not address the most extreme risks, since those risks are uninsurable, it would greatly expand the range of practically compensable harm. It would also raise the salience of liability risk for AI developers and deployers and introduce a more cautious decision maker — the insurance company underwriting the policy — into the loop. With an expanded universe of practically compensable harm, it would be more likely that the precautionary measures taken to reduce liability risk would also tend to mitigate the uninsurable risks the AI developers and deployers would still be permitted to externalize.

Finally, as indicated by the arguments in Part II of this Article, the full liability regime, with strict liability, liability insurance requirements, and expanded availability of punitive damages internalizing uninsurable risks would be preferable to any likely alternative regulatory regime. The key question in this regard is how likely such a regime is to emerge and, if it emerges at all, how many jurisdictions would adopt it. If such a regime were likely to reach

²⁰⁹ Prediction markets and other forecasting methods may be useful in reducing this uncertainty.

²¹⁰ *Car Insurance Requirements by State*, PROGRESSIVE, <https://www.progressive.com/answers/state-car-insurance-information/> (last visited Aug. 4, 2026).

²¹¹ These conversations were confidential.

widespread adoption, then it would only be worth preempting state tort law in exchange for federal legislation that included similarly robust liability provisions. By contrast, if a robust liability regime appears unlikely to emerge on the relevant timescale or to only take hold in a few small states, then it is more likely that a federal regulatory scheme could offer enough risk mitigation at low enough cost to warrant accepting preemption of state regimes. As above, explicit probabilistic forecasts, including from prediction markets, of the likelihood of different liability scenarios would be valuable for informing the debate on potential preemption.

It is also worth considering what safety-enhancing policies are likely to be on the table in any preemption scenario. If, as some have proposed,²¹² Congress were to act on a standalone preemption bill that included liability, this would be a clear setback for AI safety. If tied to broader AI regulation, preemption makes the most sense in the context of regulatory requirements that substantially burden the major AI labs, since they are also likely to be the biggest beneficiaries of liability preemption. More extreme measures like pauses and bans, however, are a poor fit for preemption, since liability relief would serve no purpose during a pause. Even after a temporary pause was lifted, it is unclear what the rationale for preempting liability would be. It is conceivable that express preemption would be included as one component of a broadly incumbent-friendly approach that establishes licensing requirements and other barriers to entry in exchange for liability relief. While licensing, independent safety testing, red teaming, and other requirements do buy some risk mitigation, the AI safety case for accepting preemption as part of this package is quite weak. A weak, incumbent-friendly regulatory regime is unlikely to offer as much risk mitigation as even the default negligence regime,²¹³ and the frontier labs are likely to be willing to support such a measure without liability protection. A key advantage of liability is that it does not require regulators to determine *ex ante* what measures are needed to make systems safe.

Regulations that might be sensibly coupled with liability preemption include a stringent pre-deployment approval process, robust monitoring and reporting requirement, safety- and incident-response planning requirements that include emergency response protocols and mandatory recalls, and third-party audits to verify compliance. The case for preemption would be

²¹² Ball & Rozenshtein, *supra* note 196.

²¹³ The reasonable care standard applied under negligence may not provide as comprehensive an inducement for developers to make their systems safe, but it would encourage them to adopt any measures that they expect a jury, with the benefit of hindsight, to deem as reasonable precautions.

strongest for prescriptive and approval regulations that carry strong penalties for non-compliance, even in the absence of substantial realized harm.

Federal law already contains the template for a preemption package that takes recourse seriously. The Price-Anderson Act has governed liability for nuclear incidents since 1957: it channels public-liability claims to the licensed operator and a federal forum, requires operators to carry the maximum available private insurance, and backs that coverage with an industry-funded pool, so that displacing the ordinary operation of state tort law came bundled with a guarantee that victims would be compensated.²¹⁴ Displacement was purchased with substitution. AI preemption proposals should be measured against that template. A package that displaces state liability while substituting a federal accountability channel of comparable strength changes the forum and funding of recourse. A package that displaces without substituting enacts the liability gap this Article has argued AI governance must close.

I further propose the following five-factor test for whether a regulatory package warrants support even if it contains a provision expressly preempting state tort law: (1) the substantive requirements buy more safety than the background negligence standard requires; or (2) penalties for violation that don't depend on actual harm are more effective in inducing compliance with the prescribed rules than negligence liability is with enforcing the ordinary standard of care; and (3) the safety gains from (1) and (2) are sufficient to compensate for the greater compliance, innovation-inhibition, and implementation costs of the regulatory regime relative to background negligence; and (4) the safety gains from (1) and (2) are sufficient to compensate for the expected safety gains from potential development of stronger liability regimes that include strict liability, liability insurance requirements, and/or punitive damages; and (5) where the package expressly preempts state tort claims, it substitutes a federal accountability channel, compensation and deterrence machinery at least as protective as the state-law liability it displaces, rather than merely extinguishing it. Quantitative estimates of the likelihood and expected risk mitigation benefits of different liability regimes would facilitate application of this test. Factor (5) operates as a necessary condition. A package that fails it fails the test regardless of its safety gains, because recourse for injured third parties is not fungible with risk reduction for everyone else: the safety gains accrue broadly, while the loss of recourse falls on the particular victims the

²¹⁴ 42 U.S.C. § 2210; *see* *Duke Power Co. v. Carolina Environmental Study Group, Inc.*, 438 U.S. 59, 63–67 (1978) (describing the Act's structure and assurance-of-compensation rationale).

displaced claims would have compensated. Under this standard, the TRUMP AMERICA draft's express disclaimer of preemption for generally applicable law, including the common law, is what satisfies factor (5).

Finally, it is worth considering the political dynamics that might lead Congress to adopt an express liability-preemption provision. That is, if the safety-oriented components of the coalition supporting the legislative package are too weak to beat back an express preemption provision, the five factors laid out above may be unlikely to hold. Obviously, once specific legislation is on the table, analysts will be in a better position to assess the risk-reduction benefits and compliance costs associated with the regulatory provisions. Nonetheless, that analysis should start from a place of skepticism that the trade is likely worth making. This is not a claim that I make from a stance of absolutism or aversion to political compromise. To the contrary, prior to joining the academy I worked for the Climate Leadership Council, in which capacity I strongly supported — and continue to support — a bipartisan compromise that exchanges a carbon fee and dividend for repeal of the Environmental Protection Agency's authority to regulate stationary-source greenhouse gas emissions. I support this policy because I believe that a carbon fee would induce greater emissions reductions at lower economic cost than the likely alternative regulatory approaches. In the AI risk context, the substantive case that liability would be redundant in the context of federal AI regulation is quite weak, given the limits of prescriptive regulation and approval regulation for ensuring AI safety laid out in part II. If that case is correct, then preemption of state liability law for AI harm is most likely to come in a context where catastrophic risk mitigation is sacrificed in favor of an approach that prioritizes innovation (possibly with an emphasis on geopolitical competition with China) and less speculative non-catastrophic AI harm like algorithmic bias, intellectual property violations, and deepfakes.

Beyond federal preemption, there has also been some effort to adopt state legislation that would offer liability protection. Most prominently, California Senate Bill 813, introduced in March 2025, would offer liability protection to AI models that are certified by private multistakeholder regulatory organizations (MROs), designated by the state Attorney General.²¹⁵ This bill built on Gillian Hadfield and Jack Clark's work on regulatory markets and Dean Ball's

²¹⁵ California SB813, LEGISCAN, <https://legiscan.com/CA/text/SB813/id/3227554> (last visited Aug. 4, 2026).

proposed refinements to tort liability.²¹⁶ In the original version, this liability protection came in the form of an affirmative defense against liability claims against MRO-certified models.²¹⁷ This was later weakened to a rebuttable presumption that the developer of an MRO-certified model exercised reasonable care.

Importantly, both versions of the bill extended this liability shield to cover lawsuits by third parties harmed by MRO-certified models, not just users.²¹⁸ In my view, this is the bill's central flaw. The market-based private governance model envisioned by SB 813's supporters may be an adequate substitute for liability in the context of user harms, where there are no externalities in the economic sense.²¹⁹ MROs would have reasonably strong incentives to set and enforce rules that protect users, so customers trust their certification and AI companies are willing to pay to get their stamp of approval. In the user harm context, where the rationales for policy intervention are asymmetric information and paternalism, these reputational incentives may well be sufficient to protect consumers.²²⁰ But consumers are unlikely to punish companies and MROs whose products generate risks for non-users. After all, AI companies will tend to prefer less onerous rules, absent market discipline.²²¹ While risks to non-users are surely somewhat correlated with risks to users, there are two reasons to think this correlation is inadequate to protect third parties. First, there are some contexts in which risks to users trade off directly with risks to non-users. For instance, developers of software for controlling autonomous vehicles must balance risks to vehicle occupants with risks to other road users, as well as tradeoffs with performance on other dimensions. Most customers are likely to prefer vehicles that protect their occupants, so MROs will have little reason to punish such choices. Second, even where user and third-party risks are correlated, the incentive for safety from user risk alone

²¹⁶ Gillian K. Hadfield & Jack Clark, *Regulatory Markets: The Future of AI Governance*, arXiv:2304.04914 [cs.AI] (2023), <https://arxiv.org/abs/2304.04914>; Dean W. Ball, *How Should AI Liability Work? (Part I)*, HYPERDIMENSIONAL (Feb. 20, 2025), <https://www.hyperdimensional.co/p/how-should-ai-liability-work-part>.

²¹⁷ See S. Judiciary Comm. Analysis, S.B. 813 (McNerney) (Cal. 2025), at 8–9, <https://sjud.senate.ca.gov/system/files/2025-04/sb-813-mcnerney-sjud-analysis.pdf>.

²¹⁸ *Id.* at 13–15.

²¹⁹ *Id.* at 2–5.

²²⁰ See Gregory Ferenstein, *California's Senate Bill 813 Is a Novel Approach to AI Safety*, REASON FOUND. (June 16, 2025), <https://reason.org/commentary/californias-senate-bill-813-is-a-novel-approach-to-ai-safety/> (explaining that “MROs must be financially independent and submit annual public reports on their evaluations and risk mitigation efforts,” supporting strong reputational incentives).

²²¹ *Id.*; see also Weil, *Don't Let AI Developers Hire Their Own Referees*, *supra* note 108 (developer-paid certification reproduces the issuer-pays conflict that discredited the credit-rating agencies, while a verifier that bears the risk it assesses faces the opposite incentive).

is typically insufficient. For instance, any single driver bears only a tiny fraction of the total social costs of their emissions—from air pollution to climate change—giving them little personal incentive to mitigate that harm. That suggests that under SB 813, the task of protecting the interests of non-users would mostly fall on the AG, in their role in designating and overseeing MROs. But the AG, or a government body assigned the task of overseeing MROs would operate under the same technical deficits and imperfectly aligned incentives as prescriptive and approval regulators. Accordingly, I think it would be a mistake to displace the tort claims of non-users as a carrot to induce AI companies to participate in a voluntary private governance system like that envisioned in California SB 813. More generally, legislators should be reluctant to trade away the recourse of third parties to tort claims in exchange for measures that do not offer them equivalent protection from risks that are imposed on them without their consent.

Given the risks of a weak federal or state regulatory regime coupled with preemption or displacement, as discussed above, it is crucial to consider the political strategy for enacting a superior, liability-centered alternative. The primary avenues for such reform lie at the state level, where tort law has traditionally been developed.²²² The most straightforward path is through the courts. As I suggest in prior work and make the case for in a forthcoming paper, state courts could be persuaded to extend the abnormally dangerous activities doctrine to cover the training and deployment of frontier AI systems.²²³ A second, though more doctrinally ambitious, judicial pathway lies in vicarious liability.²²⁴ While this route faces greater obstacles, it remains a potential complementary tool. Judicial education on AI and AI risk could play an important role in paving the way for both of these doctrinal changes.

The second path is legislative. Rather than waiting for the slow, case-by-case evolution of common law—a luxury we may not have given the speed of AI progress—state legislatures can act swiftly to clarify the availability of strict liability and punitive damages in certain cases of AI harm, and to impose liability insurance requirements on frontier model developers. Indeed, at my urging, legislators in New York and Rhode Island have taken the first step down this path by introducing bills that would establish strict developer liability in cases where an AI system

²²² See *Medtronic, Inc. v. Lohr*, 518 U.S. 470, 486 (1996) (“[T]he laws of torts are a field that has been traditionally occupied by the States.”).

²²³ Weil, *supra* note 10, at 23; Gabriel Weil, *Abnormally Dangerous Algorithms*, *supra* note 82, at 34 (proposing legislative adoption of a strict liability standard for third-party harms arising from misalignment and citing the New York and Rhode Island bills); see also RESTATEMENT (SECOND) OF TORTS § 520 (AM. L. INST. 1977).

²²⁴ See generally RESTATEMENT (THIRD) OF AGENCY (AM. L. INST. 2006).

engages in conduct that would be tortious for a human, where neither the user nor any intermediary that fine-tuned or scaffolded the model intended the conduct or was negligent with respect to it.²²⁵ Such state efforts allow for a “laboratories of democracy” approach, where a successful reform in one jurisdiction could create a model for others to follow.²²⁶

Countering the powerful and well-funded AI industry lobby requires assembling a broad and motivated coalition. The natural constituencies for such a coalition include trial lawyers and consumer protection groups, whose organizational missions align with strengthening legal protections for individuals harmed by corporate products.²²⁷ They would be joined by AI safety and public interest advocates, who can provide the crucial technical and ethical arguments for why existing negligence standards are inadequate to incentivize the mitigation of novel risks, a core theme of this Article. The insurance industry could also become a key partner, since the liability insurance requirements proposed in Part III would create a massive market for their products.

Making these reforms politically viable requires shifting the public and political conversation away from a simple “innovation versus safety” binary. The opposition will undoubtedly frame strict liability as a “tax on innovation” that will stifle progress and cede geopolitical advantage.²²⁸ An effective coalition must counter this narrative by pursuing several strategies. A primary task is to anchor the debate in concrete harms. Political momentum is more often built on tangible damages than on abstract risk analyses; this strategy mirrors the history of automotive safety, where public and congressional action was catalyzed not by statistical models alone, but by the undeniable reality of traffic fatalities.²²⁹ A high-profile “near miss” incident

²²⁵ Assemb. 8833, 2025–2026 Leg., Reg. Sess. (N.Y. 2025); H. 8052, 2026 Gen. Assemb., Jan. Sess. (R.I. 2026).

²²⁶ *New State Ice Co. v. Liebmann*, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) (“It is one of the happy incidents of the federal system that a single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country.”).

²²⁷ See, e.g., *About Us*, AM. ASS’N FOR JUST., <https://www.justice.org/about-us> (last visited Aug. 4, 2026); *About Us*, PUB. CITIZEN, <https://www.citizen.org/about/> (last visited Aug. 4, 2026).

²²⁸ See, e.g., Marc Andreessen, *Why AI Will Save the World*, ANDREESSEN HOROWITZ (June 6, 2023), <https://a16z.com/why-ai-will-save-the-world/>.

²²⁹ See JERRY L. MASHAW & DAVID L. HARFST, *THE STRUGGLE FOR AUTO SAFETY* 7–10 (1990) (describing how Ralph Nader’s *UNSAFE AT ANY SPEED* and subsequent congressional hearings focused public attention on automobile defects).

could also serve as a potent “focusing event,” the kind of crisis that creates a window of opportunity for policy change.²³⁰

Furthermore, the coalition must re-frame liability as a pro-market, pro-innovation solution. Strict liability is not central planning; it corrects a market failure, the negative externality of risk.²³¹ This narrative directly counters the claim of a “tax on innovation” by asserting that uninternalized risk is, in fact, an inefficient public subsidy. By making safety a core component of cost, liability gives developers a powerful market incentive to innovate on safety itself. Finally, these economic arguments should be paired with an appeal to fairness and responsibility. The core principle that entities should pay for the harm they cause is resonant and normatively sound.

V. Conclusion

AI governance is hampered by four closely related challenges. First, we face formidable epistemic uncertainty regarding the nature and magnitude of the risks posed by advanced AI systems. Second, there is wide social disagreement in assessments of this murky risk landscape, ranging over several orders of magnitude. Third, AI alignment and safety are unsolved technical problems. There are substantial barriers both to identifying a set of practices and procedures that would ensure the safety of AI systems and to classifying systems as adequately safe with high confidence prior to deployment. Finally, to the extent that our society does have expertise relevant to overcoming these barriers, it is highly concentrated in the private sector, primarily in the very frontier AI companies that are building the systems that may pose catastrophic risks.

These four challenges make AI governance less conducive to prescriptive and approval regulation than most governance problems. Prescriptive and approval regulation depend on a modicum of ex ante social consensus on the magnitude of the relevant risks in order to be enacted and that consensus must be at least roughly truth-tracking for the intervention to offer an acceptable balance between the conflicting goals of maximizing risk mitigation and minimizing regulatory costs and burdens. Likewise, prescriptive regulation works best in mature industries where straightforward and enforceable rules can mitigate the most important risks, and where

²³⁰ See THOMAS A. BIRKLAND, *AFTER DISASTER: AGENDA SETTING, PUBLIC POLICY, AND FOCUSING EVENTS* 1–3 (1997) (defining focusing events as sudden, rare events that draw widespread attention and create opportunities for policy change).

²³¹ See *supra* Section II.A; see also A. C. PIGOU, *THE ECONOMICS OF WELFARE* 172–96 (1932).

regulators are well-positioned to identify and enforce those rules. Approval regulation works best when the risks posed by particular products or practices are well understood, and the regulators have an understanding of those risks that is at least roughly comparable to the leading private-sector experts.²³²

Ex post liability, by contrast, is much better positioned to overcome these barriers. It does not depend on a truth-tracking consensus regarding the magnitude of AI risk. Nor does it require regulators to know what precautions are warranted or to assess the safety of specific systems ex ante. Instead, it shifts the onus to AI developers and deployers to optimize the tradeoffs between the safety and the utility of their systems. Accordingly, liability is the centerpiece of an ideal AI governance regime. Unfortunately, liability can't solve all AI risk problems, and the default liability regime falls well short of the risk-mitigation potential of an ideal liability regime. In practice, therefore, there remains substantial scope for other policies to play an important role in supporting the implementation of a robust liability regime, complementing such a regime by plugging the gaps it would leave, and substituting for the inadequacies of existing liability rules.

Three implications follow: (1) policymakers should treat strict liability as the default instrument for traceable, compensable AI harms; (2) they should invest early in low-burden 'liability infrastructure' (traceability, preservation, incident reporting) that makes ex post accountability workable; and (3) they should use approvals, prescriptive rules, and compute controls primarily as backstops or substitutes only where liability predictably fails.

The legal and political barriers to implementing a strong liability regime may tempt those worried about AI safety to trade liability away in exchange for federal AI regulation. While this might be warranted under some conditions, the analysis in part IV suggests that such trades should be treated with skepticism. Preemption of state liability rules would not only nullify the relatively weak status quo liability regime, but also cut off the potential future evolution of states' liability regime toward something closer to the ideal. While enacting a strong liability regime may seem unrealistic at present, it is not clearly less realistic than alternative policies with comparable risk mitigation potential. The same changes in the facts on the ground that might unlock the political conditions necessary for those policies could plausibly do the same for liability. While it is difficult to assess both the likelihood and the expected risk-mitigation benefits of various liability scenarios, those uncertainties cut against preemption: if federal law

²³² CASS R. SUNSTEIN, *AFTER THE RIGHTS REVOLUTION: RECONCEIVING THE REGULATORY STATE* 83–85 (1990).

displaces state tort, it should replace it with a demonstrably stronger liability regime, not a weaker substitute.