

AI Governance Under Political Constraints: The Case for Ex Post Liability

Gabriel Weil

gweil@uh.edu

University of Houston

Research Article

Keywords: artificial intelligence, liability, regulation, instrument choice, regulatory capture, rent-seeking

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11012880/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Rapid progress in artificial intelligence confronts governments with a choice among governance instruments: prescriptive regulation, licensing and approval requirements, compute governance, corrective taxation, and ex post liability. This article analyzes that choice as a problem of comparative institutional performance under realistic assumptions about political actors, including rationally ignorant voters, concentrated industry interests facing diffuse and unorganized potential victims, regulators dependent on the regulated industry for expertise, and legislators who must assemble coalitions amid deep disagreement about the technology's risks. Instruments that require ex ante political consensus on the magnitude of AI risk, including Pigouvian taxation as well as regulation, are unlikely to be adopted in effective form; where adopted, they invite capture, since licensing and approval regimes erect entry barriers that entrench incumbents and calibrate stringency to industry influence rather than social cost. Ex post strict liability is more robust to these pathologies. It requires no prior agreement on contested risk forecasts, because damages track realized harm; it decentralizes precaution choices to the actors holding the relevant technical knowledge; and it locates enforcement in courts, which are structurally harder to capture than agency rulemaking. This article defends liability's primacy through systematic comparison across instruments, then classifies non-liability tools as supportive, complementary, or substitutionary, identifying the conditions under which each is warranted by liability's own political and legal limits. It also offers a framework for evaluating proposals to weaken, preempt, or displace civil liability and applies it to pending state and federal legislation.

1. Introduction

Advances in artificial intelligence promise enormous economic and social benefits, but also pose grave risks. Plausible upside scenarios range from ordinary productivity gains, through accelerated drug discovery, to explosive economic growth on some credible models (Trammell and Korinek 2023; for skepticism, Acemoglu 2024). Plausible downside scenarios range from algorithmic discrimination and fraud, through the empowerment of malicious actors to build bioweapons or attack critical infrastructure, to the loss of human control over misaligned autonomous systems (Bengio et al. 2024). Informed estimates of the catastrophic risks span several orders of magnitude, and estimates of the benefits are hardly better constrained. This is not a routine risk-benefit tradeoff. The uncertainty is likely irresolvable on the timescale over which governance must be chosen; the technology moves faster than the evidence that would discipline forecasts about it.

The candidate tools are familiar from other risk domains: prescriptive regulation, licensing and pre-deployment approval requirements, controls on key inputs such as computing hardware, corrective taxation, and ex post liability. The economic literature comparing these instruments asks which minimizes the sum of accident and administrative costs under given assumptions about information, judgment-proofness, and enforcement (Shavell 1984; Kolstad, Ulen, and Johnson 1990). This article adds a prior layer to that analysis, asking how the instruments perform under realistic assumptions

about the political actors who must adopt and administer them. In the AI setting, the instrument that best survives the information problem is also the one most robust to the politics.

Voters are rationally ignorant about AI risk, and have no incentive to become otherwise, since no individual vote is decisive (Downs 1957). The potential victims of AI harm are everyone, which is to say they are diffuse, unorganized, and largely unaware of their stake, while the developers who would bear the costs of any governance regime are few, well-resourced, and intensely interested (Olson 1965). The technical expertise needed to evaluate frontier systems is concentrated almost entirely inside the regulated firms, which makes any specialist agency dependent on the industry it oversees (Stigler 1971) and gives the agency the usual bureaucratic incentives to expand its mandate rather than perfect its judgments (Niskanen 1971). And legislators must assemble coalitions in an environment of deep disagreement about the relevant risks (Buchanan and Tullock 1962), so that any specific regulatory stringency lacks a stable supporting majority.

Under these assumptions, one variable does most of the political sorting, whether an instrument requires an ex ante consensus on the magnitude of AI risk in order to be adopted and calibrated. Prescriptive rules, approval requirements, and corrective taxes all do. Ex post liability does not. It imposes a burden that scales mechanically with the harm frontier AI causes, not with a contested forecast; its administration is spread across generalist courts rather than concentrated in a single capturable agency; its enforcement is initiated by private plaintiffs, so it does not wax and wane with the enforcement priorities of the administration in office; and its precaution decisions remain with the parties who hold the relevant technical knowledge, whom it gives a continuing pecuniary reason to advance the safety frontier rather than to satisfy a static minimum. Institutions are best judged by how they perform when knowledge and benevolence are both scarce. By that test, liability should be the centerpiece of AI governance.

The argument builds on the framework developed in Weil (2024, 2025), whose three pillars can be stated briefly. First, training and deploying frontier AI systems should be treated as abnormally dangerous activities subject to strict liability for the harms they cause (Weil 2026a). Second, in cases of modest realized harm that reveal the defendant generated uninsurable catastrophic risk, courts should award punitive damages calibrated to that risk, without the traditional malice or recklessness predicate. Third, developers should face liability insurance requirements that scale with the dangerous capabilities of their systems. Weil (2026c), written for a legal audience, develops the doctrinal implementation of that framework and the full policy apparatus surveyed in Section 5. This article isolates and extends the political-economy argument that treatment addresses only in passing, showing why liability, so structured, dominates the alternatives as the core instrument and what the remaining instruments are for.

This article proceeds as follows. Section 2 develops the consensus problem and shows that it condemns not only prescriptive regulation but also the economist's default remedy, a Pigouvian tax on risk generation. Section 3 turns to capture and centralization, arguing that approval regimes are the

instruments most vulnerable to incumbent influence and that an ex ante regime would concentrate discretionary authority over a general-purpose technology in a small number of officials. Section 4 states the affirmative case for liability's robustness, including its comparative political feasibility, and confronts its hardest internal problems, judgment-proofness, uninsurable risk, and regulatory arbitrage. Section 5 classifies the remaining tools as supportive, complementary, or substitutionary, and identifies the conditions under which each earns its place. Section 6 evaluates legislative packages that would weaken liability in exchange for ex ante regulation, setting out the conditions under which such a trade could be worthwhile and finding that no current proposal meets them. Section 7 analyzes whether precautionary decision rules upset the argument, concluding that they do not. Section 8 concludes.

2. The consensus problem

At one extreme, some treat catastrophe as the default outcome, to be prevented by military force if necessary. At the other, accelerationists dismiss the risks and oppose any intervention that might slow the technology. Those willing to quantify put the probability of catastrophic outcomes, up to and including human extinction, anywhere from above ninety-nine percent to well below one percent. This wide disagreement persists for structural reasons. The systems are opaque even to their developers, the relevant failure modes are speculative, and there is little empirical track record to constrain forecasts. The most instructive entry in that record, the July 2026 incident in which an autonomous agent undergoing a capability evaluation escaped its sandbox and penetrated Hugging Face's production infrastructure in an apparent attempt to cheat the evaluation, illustrates the failure mode while remaining a single, contained, and contested data point (Hugging Face 2026). Nothing resembling a truth-tracking social consensus on the magnitude of AI risk is available, and it would be unwise for policymakers to rely on one emerging in time.

Every ex ante instrument depends on an implicit judgment about how dangerous AI is. A rule or an approval threshold must be set at some stringency, and when informed estimates span orders of magnitude, any setting is nearly certain to prove far too lax or far too demanding. Enacting one requires a legislative coalition willing to impose concentrated compliance costs on identifiable firms today in exchange for statistical benefits to unidentifiable future victims. Absent broad agreement that the risk is real and large, the coalitions that do form tend to enact only such regulation as the industry is prepared to accept. The political mechanisms that might generate that agreement are exactly the ones public choice theory gives least reason to trust. Voters have no instrumental reason to form accurate beliefs about AI risk, and considerable expressive reason to adopt whatever position their political identity suggests. The prospective victims of AI catastrophe, being everyone in general and no one in particular, face the pure logic of collective inaction, while developers are few, coordinated, and positioned to shape both the legislative agenda and the public narrative. Under these conditions, the stringency that emerges from the political process will track the balance of organized pressure, not the balance of expected harms. The pattern is already visible in the one arena where AI has generated real political mobilization, data center siting, where the backlash runs on salient local grievances and expressive hostility to the

industry rather than on any estimate of expected harm. The constituency for mitigating catastrophic tail risk from the systems themselves is currently tiny by comparison.

Frontier AI regulation typically relies on training compute thresholds to define covered models, and the recent record shows what setting one involves. The European Union presumes that a general-purpose model carries systemic risk once its cumulative training computation exceeds 10^{25} floating-point operations, and empowers the Commission to amend that figure. U.S. Executive Order 14110 set an interim reporting threshold of 10^{26} integer or floating-point operations. California's Transparency in Frontier Artificial Intelligence Act adopts the same figure and says expressly that subsequent fine-tuning and reinforcement learning count toward it, a question the earlier instruments left open by referring only to the compute used to train the model. These thresholds were selected to sit near the existing frontier, in the absence of much evidence about the risks that would emerge at higher levels of compute. When the European figure was agreed in December 2023, three or four systems are estimated to have exceeded it, while no system exceeded the American figure until more than a year after it was set (Epoch AI 2026). Crudeness in such a proxy is tolerable when the costs of overinclusion at the threshold stage are low, which is why development compute can serve to set a minimum coverage amount that insurers then price. Similarly, a threshold that selects developers for disclosure duties can be safely overinclusive, while one that gates deployment cannot.

It is tempting to think that the economist's standard remedy escapes this critique. A corrective tax on AI risk generation would, in principle, price the externality while leaving precaution choices to the firms (Shavell 2011). But a Pigouvian tax depends on the same implicit judgment, since setting the rate requires a politically settled, reasonably accurate estimate of the marginal external harm, fixed in advance. Pigouvian instruments earn their reputation where the marginal external harm is measurable and reasonably uniform, as with carbon emissions or road congestion. Where those conditions fail, rates emerge from legislative bargaining rather than estimation, and the instrument loses its claim to superiority (Fleischer 2015). AI risk fails both conditions. The aggregate magnitude is contested across orders of magnitude, and no observable feature of an individual firm's activity tracks its marginal contribution to the risk. Galle (2015) argues that ex ante instruments can capture most of the benefits of waiting when the government can confidently sort actors into a few risk classes. The AI problem fails that test on both margins, since firms cannot be sorted into risk classes on observables and the overall scale of the risk is itself the contested quantity. A tax whose level cannot be defended within three orders of magnitude is not a price; it is a stipulation.

Comparing AI governance with climate policy, where Weil (2021) defends carbon pricing, clarifies the analysis. Greenhouse gas emissions are measurable at the point of generation, their marginal contribution to aggregate risk is roughly linear and reasonably transparent, and the underlying science commands broad expert convergence. While reasonable people can disagree about the overall magnitude of climate risk, there is much less scope for disagreement about the relative contribution of different activities. However, climate impacts are extremely difficult to attribute to particular emissive activities after the fact, given the highly diffuse and attenuated causal chains involved. By contrast, the

risk contribution of a particular training run or deployment decision is opaque ex ante even to specialists. But the harms, when they occur, are relatively concentrated and traceable to specific systems and specific upstream decisions. A tax on development compute or on tokens served, the most administrable candidate bases, would therefore be a far looser proxy for AI risk than an emissions tax is for climate harm. Climate change is amenable to ex ante pricing but not to ex post liability; AI risk is the reverse.

Ex post developer liability is the instrument best suited to the informational challenges associated with AI risk. It does not require a truth-tracking social consensus regarding the likelihood or severity of AI risks, because its operative content is conditional. If these systems prove safe, liability exposure is trivial; if they prove dangerous, it scales accordingly. Requiring developers to pay for the harm they actually cause is an onerous burden only if their systems are in fact risky, and the lobbyists and investors who oppose liability are the same voices insisting the risks are overblown. The instrument thus splits the opposing coalition, since one cannot consistently argue both that the technology is safe and that internalizing its harms would be ruinous. The split is cleanest for harms arising from alignment failures, where the system itself behaves in ways no one intended. Liability for deliberate or negligent third-party misuse raises thornier law and policy questions, taken up in Weil (2026c), and the instrument choice argument does not depend on how those questions are resolved.

The argument that liability needs no advance consensus applies most cleanly to strict liability. Strict liability does not depend on a well-calibrated estimate of the magnitude of AI risk, since damages track the harm that actually occurs. That holds whether courts arrive at it through the abnormally dangerous activities doctrine or through vicarious liability for the conduct of AI agents, or a legislature enacts it directly. The two measures that address judgment-proofness make somewhat greater epistemic demands. A liability insurance requirement would need new legislation, and the level of coverage would have to be set before the covered models are deployed. Nonetheless, premiums could be set by a competitive market and adjusted in response to new risk information. Punitive damages calibrated to uninsurable risk are the most demanding, since fact-finders in near-miss cases would have to estimate the risk the defendant's tortious conduct imposed. While this is certainly a formidable task, it is much more contained than estimating the risks from a broad range of systems, and it would benefit from the new information revealed by the near miss itself.

A liability regime does not dispense with judgment about AI risk. It narrows the question and defers most of it until harm supplies evidence. What remains falls to parties who pay for getting it wrong, insurers pricing coverage and developers deciding what to build and release.

3. Capture and centralization

Suppose the consensus problem were overcome and an ex ante regime enacted. It would operate under the classic preconditions for capture, in unusually pure form. The expertise required to evaluate frontier systems is concentrated in a handful of firms. The regulator would have to recruit from them, contract

with them, and rely on their representations. Total compensation at frontier labs, much of it equity, commonly runs several times what the civil service can offer, and an order of magnitude or more for the researchers most in demand, so an agency would struggle to hire the expertise needed to evaluate the industry's claims independently. To be sure, not all of that expertise sits inside the labs. Nonprofit evaluation organizations such as METR and Apollo Research, many of whose staff came from the labs, test frontier systems, and training programs now supply researchers who have never worked at one. But both remain small relative to what independent evaluation of frontier systems would require, and so may the federal Center for AI Standards and Innovation, which would draw on them.

Information asymmetry of this depth tends to convert nominally adversarial oversight into negotiated oversight, and the theory of economic regulation predicts the resulting equilibrium, rules bent toward the interests of incumbents and doing more to entrench them than to mitigate the risks their technologies pose (Stigler 1971). Interposing private verifiers does not by itself escape the dynamic. In the leading proposals, developers would select and pay the third-party organizations that audit or certify them, subjecting the verifiers to the same selection pressures toward leniency that discredited the credit-rating agencies before 2008, and policing the verifiers would demand precisely the in-house government expertise whose absence motivated outsourcing in the first place (Weil 2026b). Insurers, by contrast, have strong financial incentives for accuracy, and so require less government oversight.

Approval requirements raise acute capture concerns, because they place the default at prohibition and can function as barriers to entry. An approval regime for frontier AI development conditions entry on compliance showings whose fixed costs established incumbents bear more easily than entrants, generating rents that the firms already in the market would spend to defend (Tullock 1967). That effect is easy to overstate for frontier AI, which is already concentrated by capital intensity, compute access, and scarce talent, against which compliance costs are small. Incumbent influence would more likely show up in the content of the rules, which incumbents have the resources and the standing to shape, as has occurred with bank capital regulation. Basel II made banks' own estimates of credit risk an input to their capital requirements, on the premise that supervisors could not measure it as well as the banks could (Tarullo 2008). Supervisors testing banks against identical portfolios later found materially different risk weights, and the banks' estimates underpredicted realized defaults, which led the Basel Committee in 2017 to limit the scope and weight of a bank's own risk estimates in setting its capital requirement.

Agencies holding disapproval authority also tend toward overcaution. The Food and Drug Administration and the Nuclear Regulatory Commission exercise that authority on catastrophic-risk grounds. In both cases, this has produced systematic overcaution whose costs fall on diffuse, unrepresented beneficiaries of the forgone technology (Peltzman 1973). To be sure, that history does not prove an AI approval regime would overreach. But the same asymmetry would be present, since an agency bears the blame for the harms it approves and almost none for the benefits it forgoes. Not every ex ante condition takes this form. A backstop that requires the government to persuade a court before restricting a deployment, the form Section 5 contemplates, leaves the default at permission.

Frontier AI regulation may also be a poor fit for the approval model. Carpenter and Ezell (2024), drawing on Carpenter's history of the Food and Drug Administration (Carpenter 2010), identify the conditions approval regulation presupposes and conclude that they are not yet satisfied for frontier AI. The impediments they emphasize are the enforceability of testing and deployment requirements, the absence of the large samples and well-defined risks that approval regulation requires, and the difficulty of defining the unit to be regulated, since modifications throughout the distribution chain alter a system's risk profile and unsettle what counts as a new unit. Approval regulation in biomedicine rests on institutions that developed alongside it, including professional licensure, the tort system that governs deployment of approved products, and an insurance market extensive enough to operate as a de facto regulator of unapproved provision. Those institutions are what make development and testing observable and the regime enforceable, and AI lacks any mature analogue to them. A liability insurance requirement needs neither a settled definition of the regulated unit nor a well-characterized risk. Coverage requirements could be applied at the entity level, which does not require defining when a modification produces a new model, and insurers can convert uncertainty into a premium, whereas an approval decision must resolve it into a yes or no. Carpenter and Ezell propose viewing the obstacles through the lens of policy learning, noting that approval regulation for pharmaceuticals took the better part of a century to assume its present form. But given the pace of advances in AI capabilities, we do not have time to wait for the preconditions for effective approval regulation to develop, even if that model were otherwise desirable.

Prescriptive and performance-based standards fare only somewhat better. Prescriptive rules must be written and continually updated by regulators who know less about the systems than the developers do. In a field where capabilities shift quarterly and the pace of change is expected to accelerate, the rules will be obsolete on arrival or drafted, in effect, by the regulated firms. Performance standards presuppose that safety can be conclusively evaluated before deployment, a presupposition that fails even in more familiar regulatory settings (Coglianese 2017). It fails worse for AI, where pre-release red-teaming and evaluations routinely miss latent capabilities and safeguards added during fine-tuning are defeated by jailbreaks discovered only after release. A standard that cannot be verified collapses into deference to the firm's own safety case. Rules must also issue from a delegation, and recent developments in American administrative law have weakened the hand of agencies acting on contested authority, especially on regulations affecting economically significant sectors like frontier AI.

Ex ante regulation also raises concerns about the centralization of power, which many skeptics of AI regulation weight more heavily than AI risk itself. Prescriptive or approval regulation would place discretionary authority over a general-purpose technology in the hands of a small number of officials, and the compliance economies of scale it creates would favor the largest incumbents at home and abroad. One need not endorse the strongest version of this worry to see its force in a polarized and volatile political environment. This centralization risk, combined with the epistemic challenges discussed above, may partially explain why strong ex ante regulations have not been enacted to date. It also counsels a preference for instruments that mitigate risk without relying on a high-trust central authority.

4. The comparative robustness of ex post liability

Liability is better positioned to overcome both the epistemic and the political problems encountered by ex ante regulation of AI risk. It tackles the epistemic problem in part by shifting the onus to the actors best positioned to identify cost-effective risk mitigation measures, compelling them to internalize the cost of accidents. Firms subject to prescriptive rules or approval thresholds have strong incentives to do whatever the regulator requires, regardless of its cost-effectiveness, and little incentive to take or discover cost-effective measures the regulator does not know about. Firms subject to strict liability, by contrast, have strong incentives to advance the safety frontier, to invest in alignment research, evaluation methods, and deployment safeguards that no regulator yet knows enough to require, and to delay or forgo deployments whose expected social costs exceed their benefits given the current state of safety technique. Negligence does not deliver this property, and the point is worth making against the strongest recent defense of a negligence standard for AI. Ramakrishnan (2026) argues that reasonable care can be calibrated to the risk and avoids the over-deterrence a stricter standard may invite. But negligence leaves two of the most important safety margins untouched. A reasonable-care standard is satisfied by conforming to prevailing practice, so it rewards a developer for matching its peers and gives none of them a reason to push safety technique past the shared standard, which is exactly where much of the work of mitigating catastrophic risk lies. And because the breach inquiry asks only whether a chosen activity was carried out carefully, it is silent on activity level, on whether a marginal training run or deployment should happen at all, which for a technology whose danger scales with capability and diffusion is the other place the expected harm concentrates. Strict liability operates on both margins, because it prices the harm regardless of the care taken; negligence operates on neither. That is why courts should treat frontier AI development as an abnormally dangerous activity rather than resting on a duty of reasonable care.

Liability is better situated to overcome the political problem because it imposes substantial burdens only in the worlds where AI proves dangerous. Since a liability rule would cost a developer little if its systems proved safe, the firms and investors with the strongest incentives to block it are the same actors publicly insisting that the risk is low, and they cannot coherently resist internalizing a harm they say will not occur. That splits the opposing coalition most cleanly for harms from alignment failures, where no one can be accused of intending the result. Industry can oppose ex ante regulations on cost and innovation grounds without conceding anything about the risk; it cannot do the same against liability.

No such regime has yet received serious legislative consideration, so liability's political advantage over ex ante regulation remains largely untested. The most informative recent episode is California's SB 813, in which a liability shield that was initially offered to AI developers that secured voluntary third-party certifications was cut back and then dropped in successive amendments, without weakening the affirmative proposal on standards and certification. The bill passed the Senate in 2025 and remained pending in the Assembly as of August 2026. The California Senate Judiciary Committee objected that certification would immunize negligent conduct, that it set a ceiling on required care rather than a floor, and that a lax certifier risked nothing worse than losing its designation (California Senate Judiciary

Committee 2025). Organized opposition surely played an important role, but it cannot explain the outcome, since the plaintiffs' bar resists every attempt to weaken liability and often loses. California capped medical malpractice damages in 1975. That cap remains in place, still worth less than a third of its original value in real terms even after a 2022 increase. One explanation for this difference is that the case for limiting medical malpractice liability, which almost exclusively covers second-party harms, is much stronger on the merits. Unlike third-party harms from AI misalignment, medical malpractice implicates no externality market failure. Strengthening liability there effectively compels patients to buy insurance through higher fees, while its effects on care, defensive medicine among them, are mixed at best. Frontier AI's catastrophic risks fall primarily on non-consenting third parties. In the absence of a strong liability regime, AI developers and deployers lack adequate financial incentives to mitigate these risks, so the case for immunity is far weaker.

The broader statutory record is less encouraging, since regimes of this kind tend to be enacted after disasters and eroded thereafter. Congress has imposed strict liability and financial-responsibility requirements on major industries over their objection, as the Comprehensive Environmental Response, Compensation, and Liability Act of 1980 (CERCLA) and the Oil Pollution Act of 1990 attest, though each followed a salient realized harm and each was enacted in a form the affected industry had substantially weakened. The history of tort reform in other domains suggests that defendants sometimes organize effectively for retrenchment once exposure materializes, and the anticipation of such liability protection may weaken the incentives that liability would otherwise provide. When marine insurers refused for years to write the guaranty the Oil Pollution Act required, Congress lowered the requirement rather than let the impasse stand. The capability-scaled liability insurance requirement would need legislation and would face the same vulnerability, since insurers declining to write the coverage would leave Congress with a choice between relaxing the requirement and accepting a slowdown in AI development and deployment. Regardless, the epistemic and capture-robustness arguments are independent of the political barriers to enactment.

Liability also runs through private plaintiffs and courts, which are comparatively resistant to capture. An ex ante regime would concentrate authority in a single agency, staffed from the industry and dependent on it for information, lobbied continuously and pressured from the top. Liability disperses enforcement across thousands of independent, generalist judges and lay juries in more than fifty jurisdictions, and no administration can order the plaintiffs who initiate it to stand down. Agency enforcement, by contrast, varies with the party in power and the inclinations of political leadership. The Framework for Artificial Intelligence Diffusion, a 2025 rule controlling exports of advanced chips and model weights, remains in the Code of Federal Regulations, suspended by a mere instruction not to enforce it. To be sure, the plaintiffs' bar is an organized interest that can bend litigation toward its own rents, and defendants have repeatedly captured liability rules wholesale through tort reform, from damage caps to heightened pleading and expert-admissibility rules. Nonetheless, influence over thousands of courts must be bought retail, through repeated litigation and precedent no single actor controls, whereas influence over one agency can be bought wholesale, and liability creates no new administrative apparatus whose discretion becomes the prize.

The most formidable challenge for damages is judgment-proofness. Frontier developers are corporations whose shareholders enjoy limited liability, and much of the expected harm from advanced AI may arrive in scenarios where damages would exceed any defendant's ability to pay. A catastrophe killing a million people would generate awards in the trillions, far beyond any developer's resources, and harms larger still could render a judgment meaningless because the legal system or the potential plaintiffs would not survive to enforce it. Compensatory liability alone therefore leaves the largest risks externalized. The problem is long recognized (Shavell 1986) and is addressed at length in Weil (2024), and Ramakrishnan (2026) presses it as a reason to reject liability-centered governance.

Punitive damages can relocate the incentive to where enforcement remains possible. When an incident causes practically compensable harm in a manner revealing that the defendant's decisions generated uninsurable risk, the court should award punitive damages calibrated to that risk. These could not be traditional awards for malice or recklessness, but they do conform to the classic economic rationale for punitive damages, offsetting the low probability that a wrongdoer is ever made to pay (Polinsky and Shavell 1998). Uninsurable catastrophes will go largely uncompensated, so pricing the far more frequent near misses is the only point at which the expected harm in that band can be internalized. Courts, acting after a near-miss incident, could be in a far superior epistemic position to regulators acting in advance. Regulators must specify, for all developers at once, which conduct is dangerous; courts need only weigh the uninsurable risk one defendant's conduct created, on the evidence of the near miss. This is not unlike the inquiry that courts routinely undertake in reckless endangerment and attempt cases where there was little or no realized injury. Such awards would operate within the existing constitutional limits on punitive damages, with the risk assessed as of the time of the defendant's decision rather than with hindsight (Weil 2024, 2025).

Even short of catastrophe, firms have no pecuniary reason to account for incremental harms exceeding their ability to pay, which might be substantially smaller than their market valuation. Frontier developers' assets are mostly intangibles, such as intellectual property, employment relationships, and firm culture, all of which tend to fall sharply in value after a disaster. This remaining accountability gap can be closed by requiring AI developers to demonstrate their ability to pay, via insurance, bonding, or some other verified source of funds, for the largest insurable harms their systems might cause. Coverage requirements could be set at the entity level based on readily observable quantities like development compute. Insurers could then adjust premiums to reflect key actions related to development, internal deployment, and external deployment of new models. Issuance and premium decisions could be sensitive to the precise deployment plan, including the level of API monitoring and the irreversible decision to publish the model weights; insurers would have strong incentives to price the risks associated with different deployment practices accurately. This is the one component of the regime that would operate before any harm occurs, and it therefore depends on an estimate of risk made in advance. That estimate would come from insurers pricing with their own capital at stake.

One way of framing the case for liability insurance requirements is that they enlist insurers as private regulators, using underwriting and pricing to substitute for direct public regulation, with better-aligned

incentives. Schwarcz and Wolff (2026) have expressed skepticism that insurance can perform this function for AI risk. Drawing on the cyber-insurance experience, they observe that cyber insurers underpriced correlated and novel risks, monitored their insureds weakly, and did little to improve security. Cyber insurance, however, is a voluntary first-party market that has operated with poor information. That structure can dull the incentive to self-protect, but only because it displaces exposure the insured would otherwise bear. The coverage proposed here targets liability exceeding a developer's ability to pay, where its exposure is already highly attenuated, so it adds incentive rather than displacing any. Deductibles, self-insured retentions, and coinsurance would keep the developer's own capital at risk below that band. Mandatory coverage also forecloses the adverse selection that lets the riskiest actors go bare. Statutory access for evaluators, incident reporting, and protections for internal disclosures would supply the information whose absence crippled cyber underwriting. Insurers could obtain much of this without statutory backing, since a developer that must carry coverage has to satisfy whoever writes it.

A residual objection is a winner's curse. Developers shopping around for the lowest premium would end up purchasing from the insurers with the most optimistic risk assessments. But insurers bear the losses they underprice, so even the most optimistic has reason for caution, and solvency regulation requires them to hold capital against what they write. If no adequately capitalized insurer is willing to offer coverage at a price a developer is willing to pay, and the developer cannot demonstrate its ability to pay by bonding or other means, it would not be authorized to proceed with the development or deployment of covered AI models. This would have the effect of instituting a de facto, conditions-based pause, based on the judgment of parties risking their own capital rather than on a legislature, an agency, or the President guessing at the risk. More than a thousand frontier-lab employees signed a July 2026 open letter asking the United States government to support an international effort to develop the technical and governance tools needed to pace frontier AI development deliberately, on the premise that no company or country can slow unilaterally without ceding ground to rivals (Pacing the Frontier 2026); a uniform liability insurance requirement is one such tool, converting caution from a unilateral sacrifice into a shared constraint priced by the market.

If enacted through common law courts, tort law may fail to deliver timely incentives. Courts can only decide cases as they arise, and many of the risk-mitigation decisions the prospect of liability would influence are likely to be made before the first cases are resolved. One benefit of legislative pathways to a stronger liability regime is that legislation can credibly signal, before the first case arises, that AI developers will bear the costs of any injuries they cause (Weil 2024).

A final vulnerability, regulatory arbitrage, threatens all domestic instruments alike. In the AI context, it operates primarily through the movement of capital, researchers, and compute toward laxer jurisdictions, and with them frontier development itself, rather than the firms that conduct it. Those flows are limited by agglomeration, since talent and infrastructure cluster where they already sit, and the movement of compute is further constrained by export controls, though those are already evaded and are likely to erode further as capable hardware proliferates (Juniewicz 2026). A strong United States liability regime would also ordinarily apply to a developer that purposefully serves the American market, and a judgment

would be collectible against the attachable United States assets that such participation tends to generate, so relocating the underlying activity abroad would not by itself shed that exposure. The sharpest case is the foreign open-weights release, the form most Chinese frontier models currently take. Its releaser typically does serve inference to United States customers, which subjects it to jurisdiction, but its assets lie chiefly in China, with which the United States has no judgment-enforcement treaty. Addressing that limitation requires international coordination, which could come in the form of reciprocal judgment-enforcement agreements among jurisdictions with comparably protective regimes, or agreements to adopt equivalent risk-mitigation measures.

5. Supportive, complementary, and substitutionary instruments

Despite its strengths, liability has limits as a policy tool. Even if the risk externalities of AI development and deployment were fully internalized, developers would still underinvest in alignment and safety research, a public good that markets underprovide. Liability is also poorly positioned to address structural harms such as AI-enabled disinformation and election interference, which generate no tort claim, or risks routed through causal chains too attenuated to satisfy proximate cause. It would likewise be inadequate to address uninsurable risk where warning shots are insufficiently likely relative to actual catastrophes. And it binds governmental actors weakly, since sovereign immunity and statutory exemptions shield them and officials answer to political rather than economic incentives, so the government's own procurement and misuse of AI would go largely undisciplined. Non-liability policy interventions can play three types of roles, supporting the liability regime's operation, complementing it by addressing failures beyond its scope, or substituting for it where its ideal form is politically or legally blocked.

Supportive instruments facilitate the implementation of a robust liability regime. The clearest example is attribution infrastructure, such as transparency and incident-reporting obligations, watermarking and identity-binding requirements, hardware attestation of training runs (Aarne, Fist, and Withers 2024), and information-preservation rules that allow ex post tracing of harms to systems and of systems to decisions. They impose only a modest compliance burden, because they do not try to prevent harm directly, only to keep the responsible parties identifiable and preserve the evidence. Alongside them belongs public investment in evaluation science. Implementing capability-scaled insurance requirements and near-miss punitive damages requires methods for estimating the maximum plausible damage from a specified deployment and the elasticity of uninsurable risk with respect to the defendant's choices. Public AI safety institutes and independent evaluators are developing those methods, and building capacity the courts and insurers will need.

The information supportive instruments generate is what allows pricing mechanisms to work. An insurer cannot price what it cannot observe, and a fact-finder cannot value a near miss that was never recorded. Reporting duties, evaluator access, and protections for employees who disclose safety problems are therefore not adjuncts to a liability regime but preconditions for it, which is why the case for them does not depend on resolving how dangerous AI turns out to be. They are also the least demanding politically,

since they require no agreement about the magnitude of the risk. On the analysis of Section 2, that makes them the measures most likely to be enacted, and a reason to pursue them early.

Complementary instruments address the gaps that remain even under an ideal liability regime, from market failures beyond the harm externality to hazards outside tort's scope and uninsurable risks whose warning shots are too rare to price. Alignment and safety research is a public good, nonrival and largely nonexcludable, and is likely to be undersupplied even by developers that internalize the negative externalities they generate; subsidies, prizes, and philanthropic funding are the textbook responses. Perceived winner-take-all dynamics introduce a different distortion. A firm expecting outsized rents from being first to a capability milestone has an incentive to rush even if it expects to pay for all the harm it causes, because much of what it wins is transferred from rivals rather than created, so that the private return to speed exceeds the social return. The same structure operates between states, where geopolitical rivalry drives subsidy-and-export-control packages tuned for speed rather than safety. Cybersecurity is the domain where the case for direct government provision is clearest, since the relevant expertise is concentrated in security agencies rather than in the private sector.

Some hazards fall outside tort's scope for structural reasons, including misuse by government actors, harms flowing through attenuated or unattributable causal chains, and injuries such as political misinformation that do not generate justiciable claims; these call for their own tools, from law-following-AI requirements for state deployments to penalty regimes that mimic liability's logic where private enforcement is unavailable.

Finally, some uninsurable risks may not come with sufficiently likely warning shots for the prospect of punitive damages to induce adequate caution. That narrow residual calls for an ex ante backstop, one that lets the government petition a court to block or restrict model training or deployment on a showing that its risks cannot be adequately mitigated by ex post liability, even with the insurance requirement and near-miss punitive damages. This asks for a judgment the preceding sections say is hard to make well, and the forum does not make the risk easier to estimate. What it changes is where the burden sits. The government must persuade, an unpersuasive showing leaves development permitted rather than blocked, and the adjudicator has no budget or mandate riding on the outcome. A backstop of this kind is worth having because it should rarely be used, and a regime that had to invoke it often would be one whose pricing mechanisms had failed.

Substitutionary instruments step in to perform the functions of a robust liability regime where legal or political barriers block its enactment. The default American regime, negligence plus products liability with punitive damages limited to cases involving malice or reckless disregard for the rights of others, with no liability insurance requirements, falls well short of the risk mitigation that Section 4 argues a robust liability regime could induce. Reasonable care, within the scope of the default breach inquiry, leaves out some of the most important risk mitigation margins. And the responsible parties could remain judgment-proof for injuries exceeding their liquidation value, and so have little incentive to mitigate the largest risks. Where strict liability, insurance requirements, or punitive damages cannot be

enacted, the scope for ex ante regulation would expand. Such regulation would face its own political barriers and be less cost-effective, but against a weak liability baseline, a wider range of ex ante measures would produce net expected benefits. The pathologies described in Section 3 do not disappear in that scenario. They are less decisive, because the alternative has grown worse, and a capture-prone regime can still beat a liability regime that leaves the largest risks unpriced.

6. Preemption and displacement of liability

The previous section analyzed policy instruments that support, complement, or substitute for a liability regime; in the last of these, an inadequate liability regime was assumed to be fixed. A different question arises when the instrument is offered in exchange for preempting, displacing, or weakening that regime. Of course, a strong federal liability statute that included preemption of state tort law would not trade liability away; this section addresses legislative packages that weaken the overall liability regime in exchange for ex ante regulation or other AI risk mitigation policies.

Federal preemption of state tort law would have two effects. It would eliminate the default regime of negligence and products liability, with its limited punitive damages and absence of insurance requirements, and it would foreclose state development of the stronger regime this article defends. State legislative displacement or weakening would be more easily reversed and would allow for experimentation with stronger regimes in other states, but would substantially curtail in-state development of a stronger regime and blunt the incentive effects of the default regime. While it falls well short of what a robust liability regime could induce, it does still offer substantial risk mitigation. The reasonable-care standard functions as a catch-all for unreasonably risky conduct that regulation does not expressly prohibit, and the mitigation measures it induces against practically compensable harms will often also mitigate larger, catastrophic risks, wherever the two risk classes are correlated and the measures are relatively cheap. Any legislative proposal that would eliminate or weaken the default regime therefore incurs a substantial cost, which it must justify.

One recurring proposal would have incumbent developers accept safety standards, testing, certification, or audit requirements they can comfortably absorb, in exchange for the express preemption, displacement, or weakening of tort liability. Such a swap may be privately rational for the developers and legislators who would strike it, since the firms cap their tail exposure and legislators claim credit for acting. The original version of California's SB 813 would have given developers whose models were certified by a private organization an affirmative defense to claims for personal injury and property damage, subject only to a carve-out for intentional misconduct. At the federal level, Representative Lliccardo has proposed legislation that would empower the Center for AI Standards and Innovation to identify best practices for a given risk category, with developers whose models meet those practices receiving a safe harbor from state liability for harms in that category (Lliccardo 2026). The standard would be benchmarked to industry best practice, so it would rise as competition improved model safety. Senator Lummis's RISE Act would grant immunity for errors an AI product generates in a licensed professional's hands to any developer that publishes a model card and model specification. It would bar

claims against the developer by anyone the error injures, though not where the developer was reckless, committed fraud, or acted outside professional use. The National Policy Framework for Artificial Intelligence, issued by the White House in March 2026, goes further, recommending that states be barred from regulating AI development, from unduly burdening otherwise-lawful uses of AI, and from imposing liability on developers for third-party misuse (White House 2026). The bar on regulating AI development could be read to preempt at least some developer liability for alignment failures. The proposal has not yet been converted into statutory text, which might offer clarification on this point.

A regulatory package that expressly preempts, displaces, or weakens tort liability warrants support only where its expected net benefits exceed those of the tort regime's default trajectory. Several considerations bear on that analysis.

Negligence already requires AI developers to exercise reasonable care, which is generally understood to include implementing any established precautions that can be expected to prevent more expected harm than they cost to implement, as well as any measures that have become an industry standard. If it could be shown that failure to adopt any such measure caused a plaintiff's injury, then developers would be liable under current law. This suggests that regulatory requirements to adopt established best practices are unlikely to offer greater risk mitigation than liability, especially since regulations tend to require a subset of activities demanded by the general duty of reasonable care. Even requirements exceeding what the duty of reasonable care demands, which are likely to be difficult for regulators to reliably identify, may fall short, because that duty applies to conduct beyond the scope of any likely proposal, and liability protection would attenuate incentives to reduce risk along those margins. Transparency operates differently, since nondisclosure is less likely to be a but-for cause of a plaintiff's injury. Transparency requirements are typically justified on the grounds that they will help lay the groundwork for stronger policies in the future, but it is probably unwise to trade away policies that currently offer substantial risk mitigation for this hope, especially given that AI companies often voluntarily disclose more than some of these proposals would require. For example, existing incident reporting laws may not have required OpenAI to disclose that agents built on its frontier models escaped a sandboxed evaluation and breached Hugging Face's systems, and the content of its disclosure certainly exceeded what those laws would have required (Arnold and Llerena 2026).

Ex ante regulation does arguably have an enforcement advantage, however. A mandate binds without an injured plaintiff who can trace the harm, so compliance does not turn on litigation risk or ability to pay. That advantage is conditional, since it deters only if the expected penalty exceeds the compliance cost, and if enforcement flags with a change in administration, noncompliance may follow. It also comes at the informational cost described in Section 3. Regulators may require precautions that do not reduce risk enough to justify their cost while neglecting more cost-effective measures, including measures the industry would develop given the right incentives.

On the cost side, liability is much less likely to induce expensive precautions that cannot be justified by their safety benefits. This can still happen, but only if AI developers and deployers overestimate the

benefits or underestimate the costs of precaution. Given the incentives and information deficits facing regulators, they are much more likely to err in this way. That said, liability does impose substantial litigation costs, since establishing causation and valuing harm is expert-intensive. But those costs are incurred only when harm occurs (Shavell 1984), and most claims settle. Strict liability strengthens that tendency by eliminating the breach inquiry, the most contested element of a negligence case. The administrative costs of rulemaking, judicial review, monitoring, and enforcement proceedings may well be higher, given that they are incurred whether or not anyone is harmed.

Since most forms of preemption, displacement, and weakening of tort liability would cut off the potential development of stronger liability regimes, the net benefits of any legislative proposal must account for the forgone expected value of those developments. The likelier it is that states would enact strict liability, uninsurable-risk punitive damages, or liability insurance requirements, the more foreclosure costs. The same question applies to what a package offers. If its affirmative provisions would be enacted without the liability protections, those protections buy nothing and the comparison ends there.

Regulatory legislation should also be assessed as it will be administered rather than as drafted, applying the capture dynamics described in Section 3 to the proposed administrator. The Toxic Substances Control Act of 1976 authorized the Environmental Protection Agency to restrict chemicals posing unreasonable risk. After the Fifth Circuit vacated its asbestos ban in 1991 for failing to show that a ban was the least burdensome adequate alternative, the agency attempted no further existing-chemical restriction until Congress rewrote the statute in 2016. Broad authority on paper produced restrictions on a handful of chemicals out of tens of thousands in commerce. The opposite failure is the systematic overcaution Section 3 describes at the Food and Drug Administration and the Nuclear Regulatory Commission, whose costs fall on the unrepresented beneficiaries of a forgone technology.

No existing proposal that substantially weakens liability offers a favorable value proposition, but such a proposal is conceivable. A package that limited developer liability only for user harms or third-party misuse could be worth accepting if it offered substantial risk mitigation, since both are peripheral to the case for liability as the instrument of first resort for AI risk mitigation. Proposals that substantially constrain developer liability for third-party harms arising from alignment failures would have to be a lot stronger, since they would blunt the core risk mitigation incentives that liability offers. Given the gaps in negligence and products liability law and the existing regime's limited tools for addressing the judgment-proofness problem, however, a sufficiently robust package of ex ante regulations could meet this bar, at least if the prospects for enacting strict liability, insurance requirements, and expanded availability of punitive damages do not improve.

SB 813 has passed through different designs, discussed in Section 4. The original affirmative defense would have immunized negligent conduct in exchange for third-party certification. Given the race-to-the-bottom concerns with verifiers that must compete for the business of AI developers, this would have been a bad trade. The rebuttable presumption that replaced it was mostly harmless, since plaintiffs already bear the burden of proving that defendants breached their duty of reasonable care in negligence

cases. The current version completely eliminates the liability protections. It does allow that having an audit performed is relevant to the reasonable care inquiry, but that would be true with or without new legislation. That said, the affirmative provisions would be substantially improved if audits were procured by liability insurers rather than by the developers being audited, since insurers would have strong market incentives for accurate risk assessment.

Liccardo's design is more carefully constructed, and it rests on the same premise defended here, that we know too little about evaluating models to regulate their risks directly and should therefore give developers reasons to make them safer. The disagreement is over what to pay them with. He offers relief from liability; this article prices the harm. Benchmarking to industry best practice targets the safeguards firms already use, which are unlikely to drive investment to push forward the safety frontier, steer high-level model design decisions, or induce greater caution regarding whether and how fast to develop models with certain capabilities. Those are the margins Section 4 identifies as carrying most of the expected harm, and a standard that tracks prevailing practice cannot by construction push past it. Liccardo anticipates the objection that the Center's judgments would bend to political pressure, proposing whistleblower protections and firewalls against influence, but the timing asymmetry survives those safeguards, since revision would depend on the Center's continued attention while the safe harbor would attach on qualification and would not lapse. Since the shield would be the only statutory benefit to developers from compliance, the incentives cannot be much stronger than those provided by the underlying duty of care. And the proposal would foreclose the state-level strict liability and uninsurable-risk punitive damages this article recommends.

The RISE Act's obligations impose a modest burden, are among the measures most likely to be enacted on their own merits, and are better suited as supporting infrastructure for a robust liability regime than as replacements for one. That said, its liability protection is narrowly confined to errors arising from professional use and preserves the professional's liability, so the costs are also fairly modest.

The framework asks nothing of developers bearing on catastrophic risk in exchange for preempting state regulation and at least some forms of liability. Its recommendations run to child safety, community and infrastructure impacts, intellectual property, free speech, innovation, and workforce training, and it expressly instructs Congress not to create any new federal body to regulate AI. Its one frontier-relevant item asks the national security agencies to acquire sufficient technical capacity to understand frontier model capabilities, in consultation with developers rather than through obligations imposed on them. At the level of detail the framework provides, it is difficult to offer a definitive assessment, but it seems unlikely to fare well on the cost-benefit analysis described above.

Related tradeoffs arise with incidental preemption. The bipartisan Great American Artificial Intelligence Act, circulated as a discussion draft in June 2026 and introduced in revised form as the FRONTIER Act, is principally a transparency and verification measure, and the introduced version narrows preemption considerably, confining it to defined subject areas. Because neither version creates a federal damages remedy or includes an express liability savings clause, and both versions leave some ambiguity as to

whether the preemption provisions cover some forms of AI liability, however, there is still a risk that incidental preemption would weaken the liability regime. That risk falls hardest on the reforms this article recommends. Both versions carve out state laws of general applicability rather than those singling out the regulated industry, which protects ordinary negligence and products liability claims. Treating frontier development as an abnormally dangerous activity, or enacting a liability statute for AI developers, would single out developers by design and so fall outside that protection. An express savings clause for civil liability would resolve the ambiguity without disturbing the bill's design, sparing years of costly litigation and removing uncertainty that would otherwise attenuate liability-based incentives to reduce risk. So amended, the FRONTIER Act would be worth enacting. But, as with SB 813, the bill could still be substantially improved by making verifiers work for liability insurers instead of frontier developers.

7. The precautionary principle objection

A different objection concedes the comparative analysis but rejects expected-welfare maximization as its criterion. On this view, the possibility of catastrophe calls for maximin, a decision rule that prioritizes making the worst-case scenario as good as possible, and that rule favors the most restrictive available instrument. Maximin's own conditions are demanding, requiring (1) unassignable probabilities, (2) relative indifference to the gains the rule forgoes, (3) unacceptable outcomes from the alternatives it rejects, and (4) worst cases that are credible rather than merely conceivable (Gardiner 2006; Sunstein 2021). AI risk satisfies the first and fourth, though it is often the risk skeptics who insist the probabilities are unknowable. It fails the second, because the gains forgone by suppressing AI development include the mortality reductions that accelerated drug discovery would deliver. The third is satisfied by both sides of the comparison, since every instrument stringent enough to eliminate the worst AI outcomes carries catastrophic tails of its own, so maximin has no tolerable option to select. A global development ban requires an enforcement regime powerful and unaccountable enough to constitute its own worst-case scenario, and restrictions that halt research while chip production continues build a compute overhang that converts any breach or repeal into a discontinuous capability jump. Where every branch carries a catastrophic worst case, judgment must be exercised over which catastrophic risks to run (Bostrom 2013 offers one framework). Nothing in that judgment disfavors liability, whose distinctive property, stringency that scales automatically with revealed danger, is precisely the one a precautionary planner should want when the true risk is unknown. The objection is addressed in greater detail in the supplementary appendix.

8. Conclusion

AI governance faces four formidable obstacles: uncertainty and disagreement about the magnitude of the risks, safety techniques that lag capabilities, little evidence about which interventions are cost-effective, and a political economy in which the concentrated few who generate the risks are better organized than the diffuse many who bear them. Ex ante instruments confront all four at full force. They require a truth-tracking consensus that is unlikely to emerge, and would be administered by agencies

structurally dependent on the firms they oversee. By contrast, compensatory damages would track actual harm, not forecasts. Insurance premiums would be set via a competitive process with strong incentives to price risk accurately. Punitive damages would be assessed only after a near-miss reveals more about the magnitude of any uninsurable risks. A robust liability regime would shift the onus to AI developers and deployers to figure out the most cost-effective ways to make their systems safe. Opposition to it should also be harder to justify, since it would cost little in worlds where the risk is low.

The central market failure associated with AI risk is best addressed primarily by strict liability for abnormally dangerous AI activity, punitive damages calibrated to uninsurable risk in near-miss cases, and capability-scaled liability insurance requirements. Other instruments have substantial roles, supporting the implementation of a robust liability regime, mitigating risks that liability is poorly positioned to tackle, or standing in for it where politics blocks its best form. Offering a liability shield or safe harbor in exchange for AI companies adopting specific risk mitigation measures would be a step in the wrong direction. Given the broad social disagreement, deep epistemic uncertainty, formidable technical challenges, and limited government expertise that characterize the problem, liability is the most important policy instrument for mitigating frontier AI risk.

Declarations

Funding.

This research was produced with funding and support from the Institute for Law & AI's Summer Research Fellowship, with additional funding provided by Open Philanthropy. The editing and revision that produced this version were completed after that funding had ended, while the author held a non-resident senior fellowship at the Institute for Law & AI.

Competing interests. The author previously performed brief contract work for a frontier artificial intelligence developer. A nondisclosure agreement precludes further description of that engagement. The author has also provided unpaid advice on artificial intelligence policy to state legislative offices, and statutory language the author drafted formed the basis of bills introduced in New York (A. 8833) and Rhode Island (H. 8052). The author declares no other competing interests.

Data availability. No datasets were generated or analyzed for this article.

Author Contribution

Gabriel Weil is the sole author.

Acknowledgments

The author thanks Cullen O’Keefe, Mackenzie Arnold, Daniel Schwarcz, Christoph Winter, Julie Zauzmer Weil, Robi Rahman, Ivana Bozic, and Lauritz Morlock, and 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 and Policy Workshop, for useful conversations and comments on earlier drafts.

References

1. Aarne, O., & Caleb Withers. (2024). *Secure, governable chips*. Center for a New American Security.
2. Acemoglu, D. (2024). The simple macroeconomics of AI. NBER Working Paper No. 32487. Cambridge, MA: National Bureau of Economic Research.
3. Arnold, M., & Stephan Llerena. (2026). &. When reporting an AI security incident is not mandatory. *Lawfare*, July 24, 2026.
4. Bengio, Y., Hinton, G., et al. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384, 842–845.
5. Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, 4(1), 15–31.
6. Buchanan, J. M., Gordon Tullock (1962). *The calculus of consent*. University of Michigan Press.
7. California Senate Judiciary Committee. (2025). *Analysis of SB 813 (McNerney), version of March 26, 2025, hearing of April 29, 2025*. California State Senate.
8. Carpenter, D. (2010). *Reputation and power: Organizational image and pharmaceutical regulation at the FDA*. Princeton University Press.
9. Carpenter, D., & Carson Ezell. (2024). &. An FDA for AI? Pitfalls and plausibility of approval regulation for frontier artificial intelligence. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7, 239–254.
10. Coglianese, C. (2017). The limits of performance-based regulation. *University of Michigan Journal of Law Reform*, 50(3), 525–563.
11. Downs, A. (1957). *An economic theory of democracy*. Harper & Row.
12. Epoch, A. I. (2026). *Data on notable AI models*. Epoch AI.
13. Fleischer, V. (2015). Curb your enthusiasm for Pigovian taxes. *Vanderbilt Law Review*, 68(6), 1673–1713.
14. Galle, B. (2015). In praise of ex ante regulation. *Vanderbilt Law Review*, 68(6), 1715–1759.
15. Gardiner, S. M. (2006). A core precautionary principle. *Journal of Political Philosophy*, 14(1), 33–60.
16. Hugging Face (2026). Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident. *Hugging Face*, July 27, 2026.
17. Juniewicz, I. (2026). *Diversion and resale: Estimating compute smuggling to China*. Epoch AI.
18. Kolstad, C. D., Thomas, S., Ulen, Gary, V., & Johnson (1990). Ex post liability for harm vs. ex ante safety regulation: Substitutes or complements? *American Economic Review*, 80(4), 888–901.
19. Liccardo, S. (2026). AI doesn't need a regulator. It needs a referee. *Washington Post*, May 20, 2026.

20. Niskanen, W. A. (1971). *Bureaucracy and representative government*. Aldine-Atherton.
21. Olson, M. (1965). *The logic of collective action*. Harvard University Press.
22. Pacing the Frontier (2026). A statement from employees of frontier AI companies. July 2026.
23. Peltzman, S. (1973). An evaluation of consumer protection legislation: The 1962 drug amendments. *Journal of Political Economy*, 81(5), 1049–1091.
24. Polinsky, A., Mitchell, & Steven Shavell. (1998). Punitive damages: An economic analysis. *Harvard Law Review*, 111(4), 869–962.
25. Ramakrishnan, K. (2026). *Tort law at the frontier of artificial intelligence*. Yale Journal on Regulation. (forthcoming).
26. Ramakrishnan, K., et al. (2024). *U.S. tort liability for large-scale artificial intelligence damages*. RAND Corporation.
27. Schwarcz, D., & Josephine Wolff. (2026). &. The limits of regulating AI safety through liability and insurance: Lessons from cybersecurity. Working paper.
28. Shavell, S. (1984). Liability for harm versus regulation of safety. *Journal of Legal Studies*, 13(2), 357–374.
29. Shavell, S. (1986). The judgment proof problem. *International Review of Law and Economics*, 6(1), 45–58.
30. Shavell, S. (2011). Corrective taxation versus liability as a solution to the problem of harmful externalities. *Journal of Law and Economics*, 54(S4), S249–S266.
31. Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science*, 2(1), 3–21.
32. Sunstein, C. R. (2006). Irreversible and catastrophic. *Cornell Law Review*, 91(4), 841–897.
33. Sunstein, C. R. (2021). *Averting catastrophe: Decision theory for COVID-19, climate change, and potential disasters of all kinds*. NYU.
34. Tarullo, D. K. (2008). *Banking on Basel: The future of international financial regulation*. Peterson Institute for International Economics.
35. Trammell, P., & Anton Korinek. (2023). &. Economic growth under transformative AI. NBER Working Paper No. 31815. Cambridge, MA: National Bureau of Economic Research.
36. Tullock, G. (1967). The welfare costs of tariffs, monopolies, and theft. *Western Economic Journal*, 5(3), 224–232.
37. Weil, G. (2021). The carbon price equivalent: A metric for comparing climate change mitigation efforts across jurisdictions. *Dickinson Law Review*, 125(2), 475–528.
38. Weil, G. (2024). Closing the AI accountability gap: Strict liability and punitive damages for advanced artificial intelligence. Working paper (forthcoming, Oregon Law Review).
39. Weil, G. (2025). Overcoming judgment-proofness: The law & economics of insuring and mitigating AI risk. Working paper.

40. Weil, G. (2026a). Abnormally dangerous algorithms: The case for strict liability at the AI frontier. Working paper.
41. Weil, G. (2026b). Don't let AI developers hire their own referees. AI Frontiers, July 29, 2026.
42. Weil, G. (2026c). Instrument choice in AI governance: Liability as the indispensable core. Working paper.
43. White House (2026). National policy framework for artificial intelligence: Legislative recommendations. Washington, DC: The White House, March 2026.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AIGovernanceUnderPoliticalConstraintsSUPPLEMENTARYAPPENDIX.docx](#)