Self-Regulation at the Frontier: Structural Limits of AI Safety Frameworks under SB 53 and Beyond

The Economy Research Editorial[1,2]

[1]The Economy Research, 71 Lower Baggot Street, Dublin 2, Co. Dublin, D02 P593, Ireland

[2]Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

The recent introduction of state-level "frontier AI" laws in the U.S. is testimony to the increasingly wide gulf between the rate of technological acceleration and the capacity of the U.S. Government to respond. SB 53 in California, for instance, to go into effect in 2026, is a case of regulatory design that envisions transparency and internal firm safety procedures as the principal method for ensuring private innovation operates at an appropriate level of public safety. This paper argues that this vision is incomplete. Transparency is an important component of accountability, but it provides neither enforcement nor a solution to the underlying agency problem embedded in firm-designed risk profiles, evaluations and commitments under market competition. From 2023-2026, frontier AI markets are at an unusually critical juncture of concentrated capital, diffuse capability, and significant geoeconomic attention. In such a moment, it is predictable that self-crafted safety regimes tend toward minimal viability instead of sustainable public protection. The paper presents structural critiques of SB 53's design, as well as the publicized safety mechanisms of the main frontier developers, to make the case that effective safety necessitates the institutional separation of standard-setting, evaluation, and deployment authorities, complemented by independent verification capacities and robust penalty regimes.

1 Introduction - Transparency Without Enforcement

The standard American narrative of AI governance has become increasingly unconvincing: it's possible to preserve innovation while managing safety through a patchwork of self-commitments, bottom-up enforcement, and market discipline. This is not a problem of transparency not mattering or private ordering playing no role. Rather, the problem is that the received framing confuses transparency with enforceability. It assumes that public access to corporate procedures, in the form of prior disclosure and mid-course reporting, will automatically appear as control over corporate behavior. But transparency exists as an informational state, not a causal state: it is neither enough in itself nor always effective without either medium,term enforcement tools such as verification capacity and legal sanctions or long-term incentives such as stable, costly commitments that are more valuable than the benefits of fast deployment.

SB 53, signed by California Governor Gavin Newsom, effective 1 January 2026, should be viewed as a best-case example of trying to outdo voluntary language with formal obligations, without building a full command and control regime.[1] It establishes a legal category of "large developers" of "frontier models," mandates the maintenance of "frontier AI safety frameworks," and insists on periodic, publicly verifiable disclosures.[2] In effect, it prescribes "disclosure first, then accountability, then enforcement." SB 53 is not an inadvertent step toward regulation, but an intentional sequencing.[3] The 2023 executive order emphasized risk management, reporting, and internal governance processes, but it did not create a comprehensive licensing or external approval regime.[4]

In the meantime, even as the EU moved toward a regulation premised on a binding, risk-based, prescriptive legal regime[5], the U.S. delayed, saying the same thing through non-binding norms and standards. The EU AI Act entered into force in 2024 and began phased application in 2025, reflecting a more prescriptive and externally defined model of governance. The U.S. is unhurried and incremental, while the EU is prescriptive and risk-focused. This institutional difference reveals contrasting assumptions about the production of safety, either through self-commitments to go internal or through external, legal coercion.

SB 53 suffers from the drawbacks of premature self-evidentiary leverage. It is a uniquely American regulatory ecosystem: aspirational in scope, cautious in coerciveness. SB 53 establishes reporting and framework obligations for large developers, but it does not prescribe the substantive content of acceptable safety practice in a way that would sharply constrain engineering or deployment decisions.[6] It's hard to trust a rule that makes compliance visible and easier to regulate, without also allowing the regulator to enforce a large number of widely dispersed operations.

The upshot is that SB 53 should be interpreted as a transparency statute for frontier AI development, not a comprehensive command and control framework. It makes the self-directed commitments of frontier AI more visible without meaningfully constraining them.[7] At most, it improves safety where disclosure is relatively cheap and where corporate and public incentives already overlap; the harder problems are independent verification and the correction of adverse incentives.

2 Principal-Agent Problem

The structure that SB 53 rests upon is standard within the field of regulatory economics - a principal sets an agent up with safety-relevant discretion but misaligned incentives. The principal here is the public, functioning as a state authority; the agent is the frontier developer, functioning through its officers and investors. The moral hazard is not simply a matter of firms wanting to profit, but that the regulated party is the only entity with sufficient technical expertise to identify relevant risk benchmarks, develop assessment methods, and determine practical ways to reduce risk, given the speed at which frontier development happens. The regulator's dependency is not an accident of circumstances, but fundamental to the situation.

The principal-agent problem is exacerbated when internal governance structures, rather than regulations themselves, become the central mechanism of compliance. While the statute mandates that "large developers" create and maintain a frontier safety framework, the law does not, by itself, dictate the specific risk tolerance of such a framework. It thus comes about that the unit of compliance is a self-authored document, the content of which is driven by firm incentives and initially managed by the firm itself.[8]

This predictably produces the following governance dynamic: the regulation institutes a new bureaucratic document (the safety framework), which the regulated organization will then manage to achieve maximum legal credibility while retaining maximum operational latitude. When legal obligations focus on the possession of a framework and compliance with that framework, the dominant strategy will be for firms to make their frameworks believable to outsiders while allowing themselves the maximum freedom possible.

The more that the framework can be presented as "credible" without obligating the firm to any truly hard constraints, the more likely it is that the framework will become an obstacle rather than a stopper. While the transparency mandate in SB 53 may still seem appealing, transparency can only constrain behavior indirectly, through reputation mechanisms, and by informing enforcement actions. However, reputation can only go so far when it comes to complex systems, with opaque workings and planned disclosure, are concerned. Evaluation results from frontier AI development may depend on proprietary information on models, test conditions that cannot be publicly disclosed for security reasons, or the results of proprietary red-teaming that may weaken security if shared widely. In short, firm discretion remains where it is most critical, as a matter of choosing what to make publicly readable, what to keep private, and what to present to trusted auditors or regulators.

According to a report from the Office of Governor Newsom, California's SB 53 sets new requirements for developers of advanced artificial intelligence, moving beyond cosmetic accountability measures and seeking to ensure that company processes and controls are actually sufficient to limit risks in line with public expectations. The only relevant question is not the existence of a framework, but the suitability of that framework's risk thresholds, evaluative benchmarks, and mitigations for the public risks to be shouldered.

2.1 The Self-Regulatory Model of AI Governance

The self-regulatory model built into SB 53 can be seen not as an instance of substantive, but rather of process, regulation. The firms set the key variables: which capabilities are relevant; what metrics, or tests, measure these capabilities; which thresholds constitute unacceptable risk; and what mitigations are needed before a given step

can be repeated. The state's role is largely to ensure that the apparatus exists, is publicly disclosed in specified ways, and is followed as written.

This model has the potential to converge toward an equilibrium through both competition and public oversight. If one company has drastically stricter rules than others, it may lose talent, capital, and market share. If one has significantly weaker rules than its competitors, it may face negative reputation impacts, consumer backlash, and potentially political intervention. Eventually, the argument goes, firms will be nudged toward a position on the continuum of safety-criticality that is both competitive and maintains legitimacy.

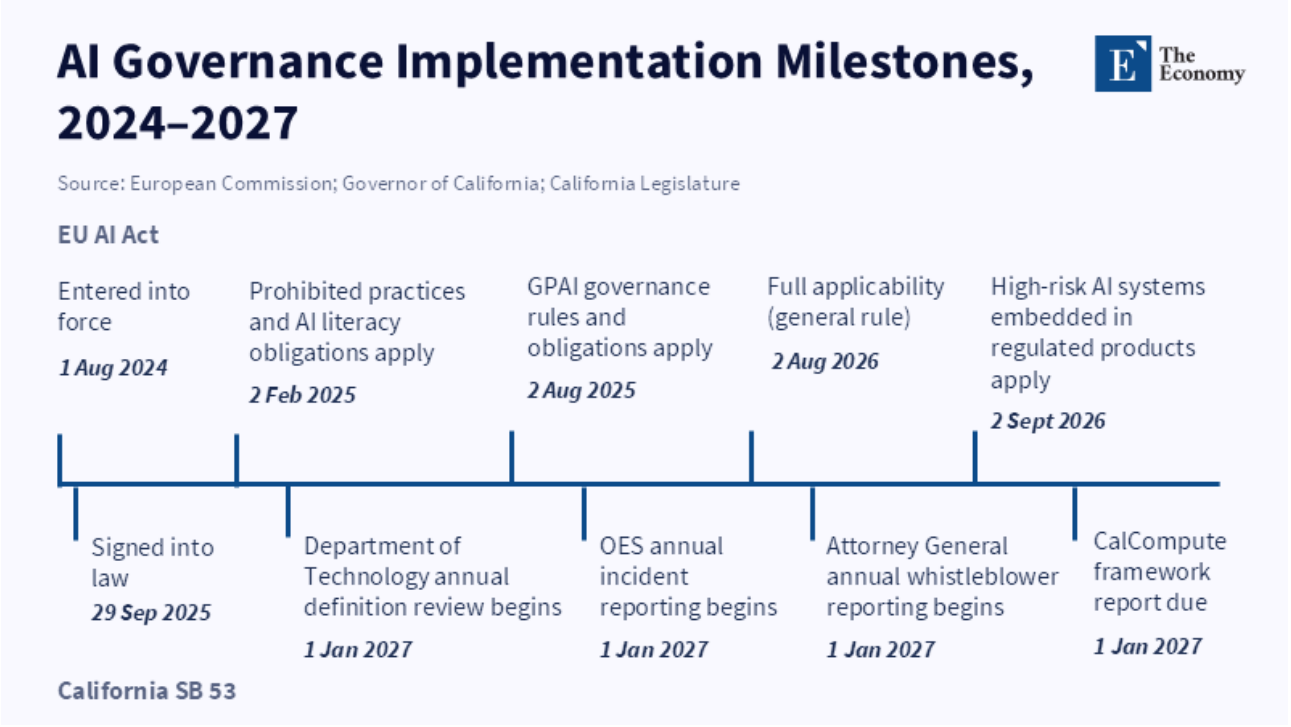

Figure 1: The policy timeline shows that the EU moved earlier toward externally defined, phased obligations, while SB 53 remains centered on disclosure, reporting, and later institutional review.

The major challenge to this model is that frontier AI is not a mature industry where product development cycles are slow and consumer demands are stable. Rather, it is a swiftly evolving field where the advantage conferred by the latest marginal increases in capability can be discontinuous and where the first-mover advantage is enormous. Evidence from the Stanford AI Index on the speed at which adoption by organizations and private investment is accelerating in the mid-2020s illustrates the extreme competitive pressures to deploy rapidly and widely.[8] In those environments, any eventual equilibrium, if one exists, may be too late to prevent the kinds of harms that drive the need for these frontier safety statutes.

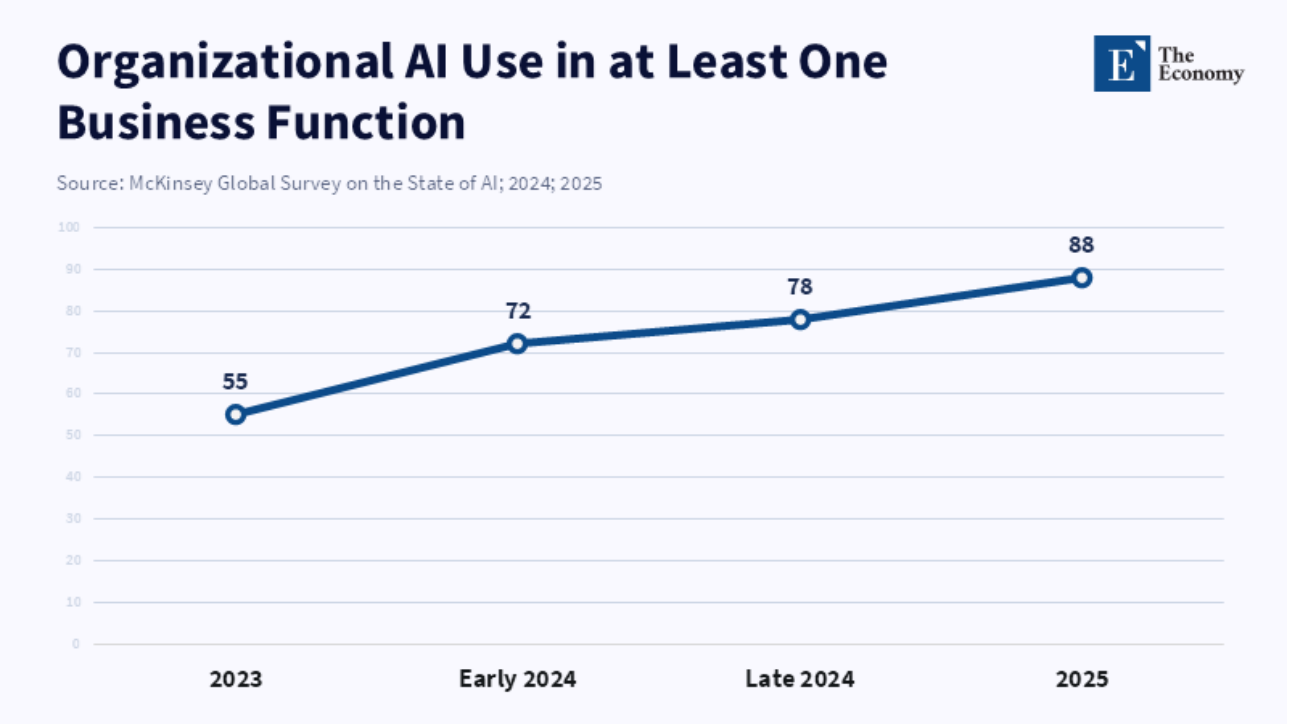

Figure 2: Rapid growth in organizational AI use shows why governance based on slow convergence and voluntary adjustment may be overtaken by deployment speed.

Furthermore, the mechanism of "public scrutiny" is inherently limited by information asymmetry. The public is not able to observe or evaluate what went on inside the lab (as opposed to just its outputs). According to Mintz, regulators face difficulties in telling apart between genuinely conservative security limits and targets that are easily achieved through clever framing. As a result, effective self-regulation requires oversight that goes beyond relying on good faith, especially since laws like California's SB 53 give the Attorney General authority to periodically update key definitions to better address advances in AI development. Effective oversight, therefore, requires externally auditable standards and quantifiable commitments, not transparency alone.

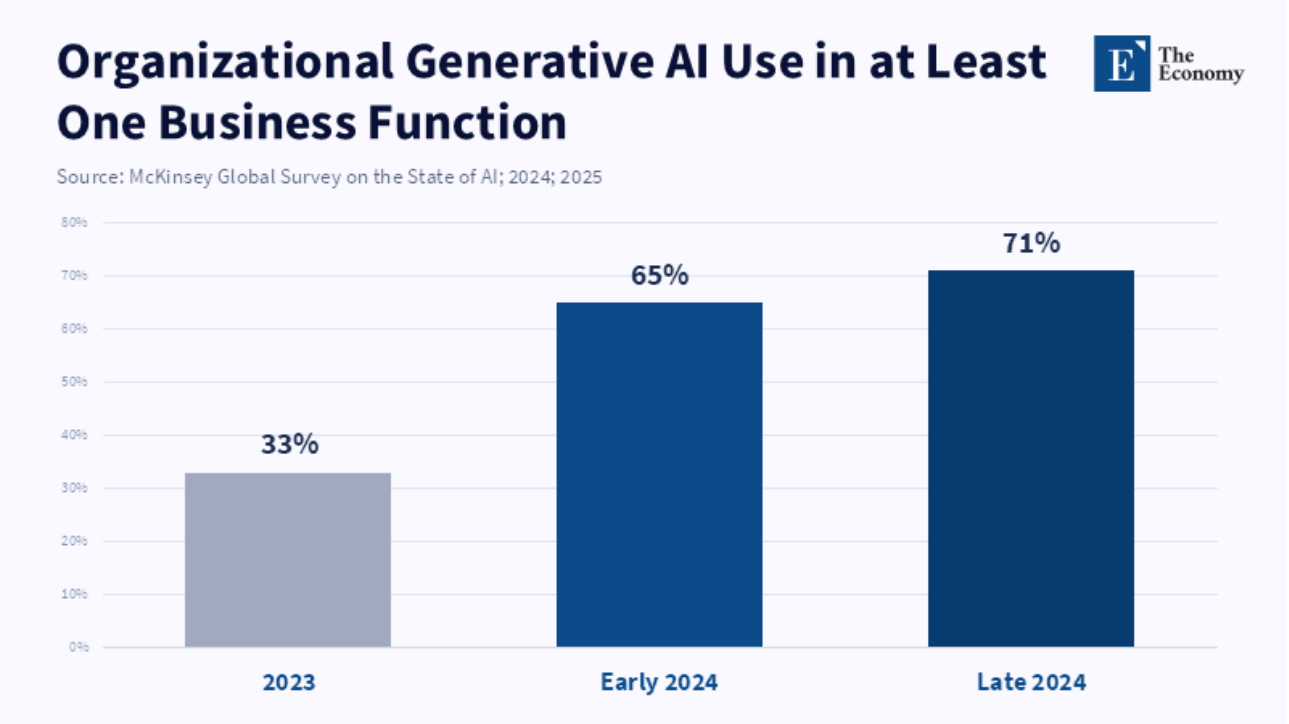

Figure 3: The particularly steep rise in generative AI use reinforces the argument that the market is scaling faster than self-regulatory oversight can mature.

The OECD's research on managing AI risk is consistent with this, as it implies that a broad array of internal governance structures, incident reporting procedures, and transparency methods exist within companies, although information regarding core inputs such as training data practices differs significantly among organizations.[10] This is not a simple technical detail. It signifies that "transparency" is not one switch to flip on; it is a deeply contested domain where companies selectively disclose in areas where disclosure is inexpensive and limit disclosure in areas where it is expensive. A regime that puts the spotlight on disclosure can never escape the problem of assuming away the incentives that create non-disclosure.

2.2 Structural Incentive Conflict in Frontier AI Safety

The inherent incentive conflict of SB 53 is clear: the entity that has the greatest to benefit from accelerating deployment is granted significant power over determining what "safe enough" is. This is not an issue to be dismissed lightly. It is, quite clearly, the critical vulnerability of the regime.

The conflict is amplified by the nature of risks posed by frontier AI. These harms- particularly those termed "catastrophic" or "critical"-have an extremely low probability but high impact and are associated with extreme levels of uncertainty. There is no shared empirical reference point in this space that can force convergence. Unlike food safety laws, which are built on many years of established scientific data, the development of rules for frontier AI safety still depends on debated models, shifting scenarios, and evolving benchmarks. According to a recent statement from the office of Governor Newsom, California's new SB 53 introduces strengthened requirements for frontier AI developers, reflecting how rapidly these standards can change depending on who sets the benchmarks.

As noted in NIST's AI risk management framework, the nature of AI is inherently socio-technical and risks

derive from complex connections between technical factors and social contexts, implying that the setting of metrics and thresholds necessitates judgment.[11] This is particularly true when dealing with frontier models. There is no objectively definable line between acceptable and unacceptable risk. If the regulated entity controls how that judgment is exercised, it effectively controls the boundary between risk and compliance.

In a typical principal-agent context, the solution for such conflicts is often the separation of functions. The agent is tasked with implementing certain actions, while a principal (or an independent entity) sets evaluable standards, confirms performance and imposes credible penalties. The law compensates the public with openness measures, including a new mandate that developers notify the California Office of Emergency Services if such behavior is detected. The dilemma of SB 53 is not one of opacity versus creativity, but of legibility versus independence. The question for policymakers is whether legibility alone is sufficient to discipline an industry whose core incentives reward rapid deployment and continual reinterpretation of its own commitments.

3 Empirical Evidence

The empirical case for skepticism about self-regulatory frontier safety regimes doesn't need to presuppose malevolence. It hinges instead on observing how companies operate when commitments collide with competitive advantage, profits, or international demand. Over the past half-dozen years, the frameworks published by major frontier developers have gotten increasingly elaborate and are frequently framed as signs of industry maturity in the mid-2020s. What is far more salient, however, are the ways in which these schemes are constructed to retain discretion through high harm thresholds, flexibility in interpretation, selective scoping, and an explicit possibility of revision.

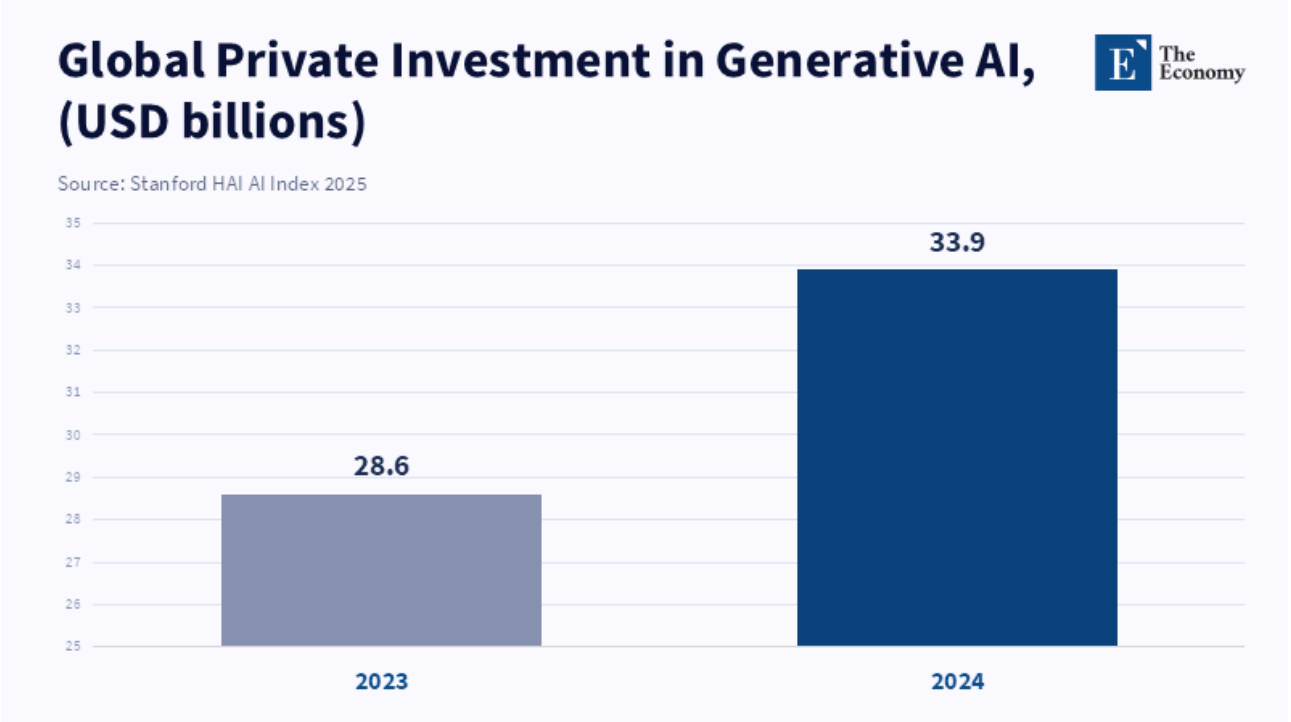

Figure 4: Rising private investment in generative AI helps explain why firms face persistent incentives to preserve discretion rather than accept hard external constraints.

The trend is more than theoretical. The frequency of reported media incidents and hazards relating to AI has

materially increased through the 2020s; according to OECD, the average monthly frequency of media-reported AI incidents and hazards went from 92 in 2022 to 324 in 2025, although the portion these make up of overall media coverage of AI has shifted.[12] While media reporting is hardly a flawless proxy for underlying harm, the trend signals that the negative externalities of AI development and deployment are becoming both more visible and more politically salient at precisely the moment that frontier capability races are accelerating.

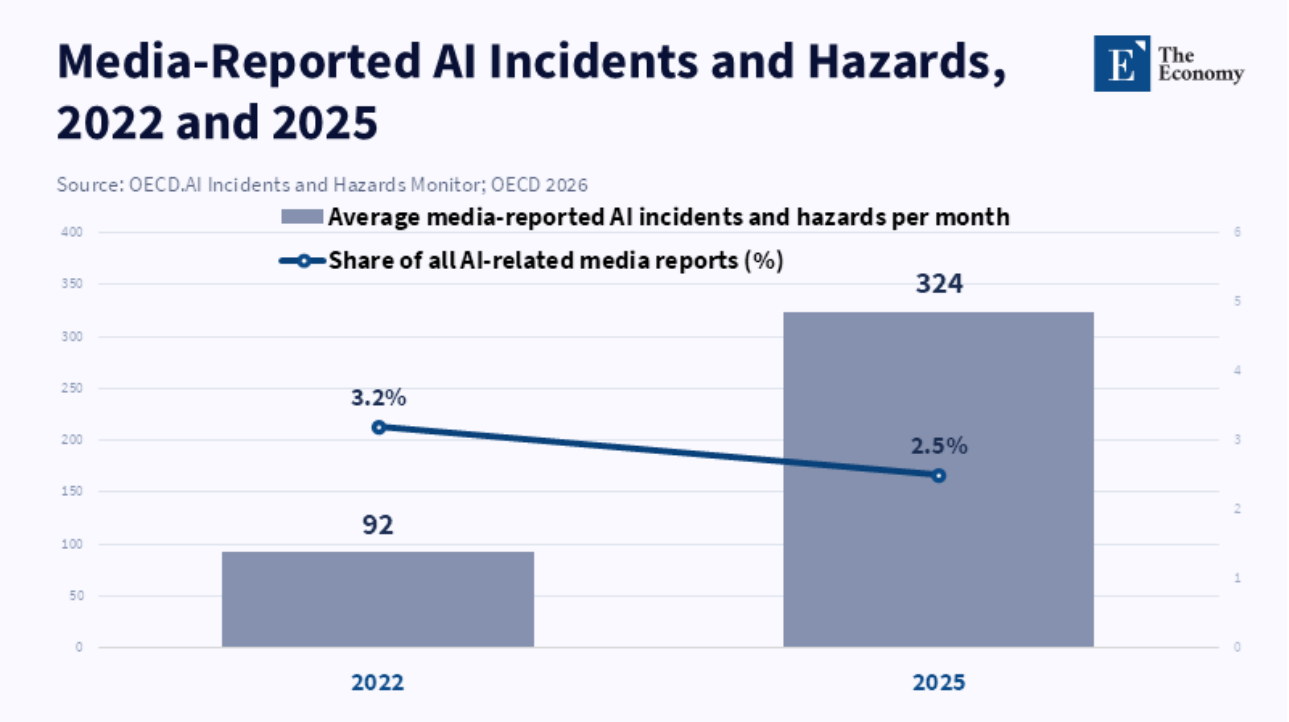

Figure 5: Media-reported AI incidents rose sharply even as AI coverage broadened overall, suggesting that visible harms are growing alongside wider deployment.

International institutions, meanwhile, have noticed a boom in corporate ”frontier safety frameworks.” The International AI Safety Report 2026 remarks that in 2025, twelve different companies had published or updated such frameworks to the dissemination of governance templates, but perhaps more to the political benefits of appearing ”responsible.”[13] The key policy problem becomes, do the frameworks represent substantive constraints, or are they essentially management tools to maintain oversight scrutiny?

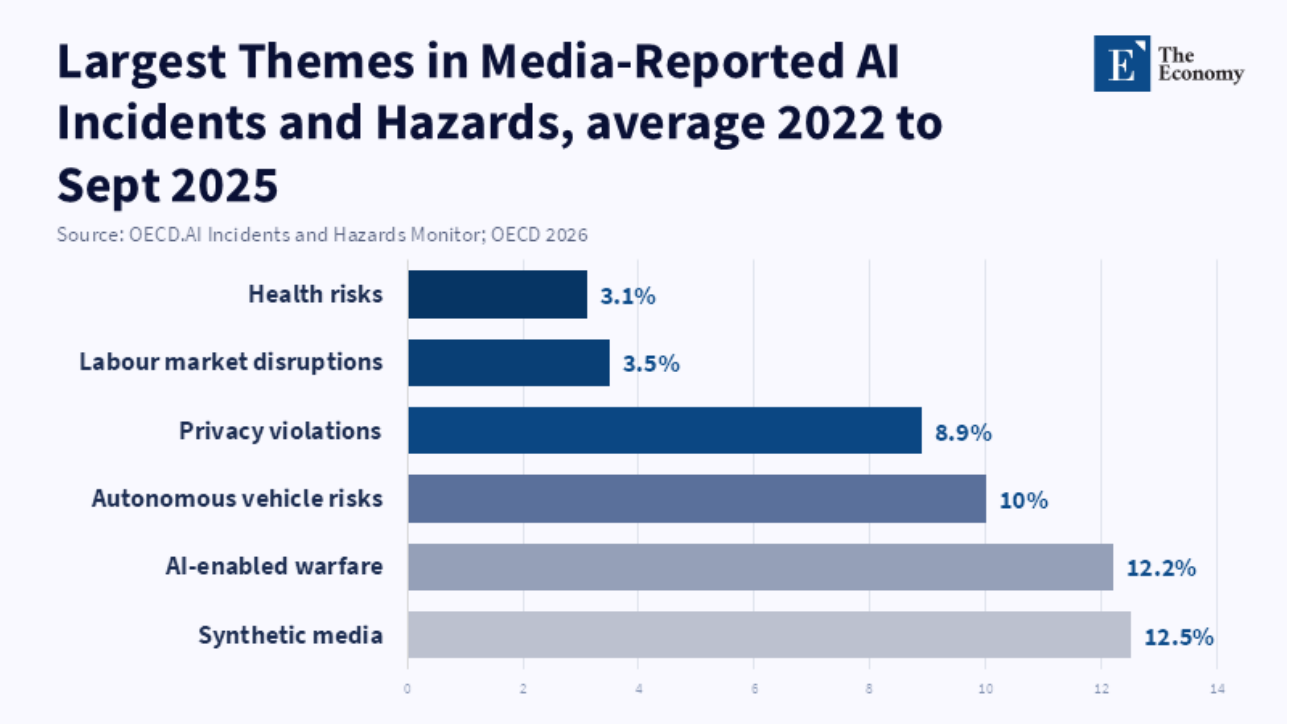

Figure 6: This is the point where the text moves from broad patterns to case studies. A thematic breakdown of incidents works well as the bridge.

The following case analysis draws on the companies listed in the prompt – OpenAI,[14] Google DeepMind,[15] xAI,[16] and Anthropic[17] – not to resolve every factual dispute surrounding them, but to reveal recurring structural mechanisms of ambiguity utilization, selective application, shift of evaluative domain, and programmatic retreat.

OpenAI's Preparedness Framework sets forth a formal safety regime whilst explicitly carving out significant space for discretion by establishing "severe harm" at a very high level of abstractness – "death or grave injury of thousands of people, or hundreds of billions of dollars of economic damage" – and openly declaring that many risks will not reach this threshold.[18] A high threshold has abstract defensive rationales, concentrating governance on the very worst outcomes. In practice, it crucially determines how the firm is required by its own framework to treat these situations as risks worth stopping the organization for. When a firm defines the boundary of "severe harm," it decides which safety failures can reasonably be framed as existential, and which simply represent manageable externalities.

The mismatch can become glaring if this threshold is compared with SB 53's catastrophe framing: "50 or more deaths or serious injuries, or property damage exceeding one billion dollars".[19] This difference is not trivial. When an organization defines its risk-management framework according to a much higher harm threshold than a public statute envisions, "operating according to internal policy" may become an excuse for legal compliance coupled with substantial practical policy failure relative to public standards.

OpenAI's framework also reveals how staged development stages and internal governance procedures promote a high degree of flexibility. The framework specifies various categories of tracked risks, from biological and chemical to AI self-improvement; these categories correlate with different capabilities thresholds and corresponding requirements for safety measures. Internal groups are charged with assessing the adequacy of safeguards.[20]

While this can be legitimate, the judgments themselves are internally derived; the framework stresses that it focuses on future capabilities outperforming current models. The danger of a moving frontier like this is that the strictest standards are reserved for hypothetical futures, while the present is governed by relatively loose guidelines and practices that depend on internal discretion.

Google DeepMind's Frontier Safety Framework reveals a different mechanism: evaluation can be narrowed through internally defined judgments about whether new capabilities are 'meaningful' or 'material' enough to trigger further scrutiny.[21] While logical and practical on its face (repeated detailed reviews are a huge administrative burden), this creates an interpretative hinge. Whether any particular new capability development can be considered "meaningful" or "material" is left to the developer's discretion. According to California Senate Bill 53, large frontier AI developers are required to create and publish a framework that defines and assesses thresholds for catastrophic risks and details their risk reduction methods. This process entails setting specific thresholds and mechanisms to determine when and how risks are evaluated or addressed, which may limit the scope of evaluation or justify not expanding it.

According to an article by Francesca Biagini and colleagues, some systemic risk frameworks recommend allocating risk randomly to individual institutions [22] before aggregating their overall risks, which can help identify system-wide vulnerabilities without imposing enforceability requirements at the level of each institution. This distinction is key; a framework that sets enforceable criteria for external misuse but not for internalized systemic risks will, under pressure, always tilt toward mitigating external, concrete dangers and against the more difficult, destabilizing systemic threats.

XAI's published risk management framework displays how concrete numbers may still be constructed to yield maximal discretion. The framework sets explicit numerical criteria, such as requiring the dishonesty rate to be below one out of two on its dishonesty benchmark, MASK, while assuring that additional thresholds will be developed.[23] It also makes reference to limiting specific query reply types (e.g., bio/chem answers) and states that these targets are provisional and subject to adjustment as evaluations evolve.[24] This implies two things in policy terms: firstly, thresholds set at values where substantial failure is still accepted (here, a dishonesty rate nearing fifty percent) likely do not correspond to the public's expectations of what is safe enough from systems meant to handle critical informational functions; secondly, and perhaps more crucially, the ability to make explicit the non-binding status of these standards as asserting they are provisional and determined internally gives the company full discretion over their interpretation.

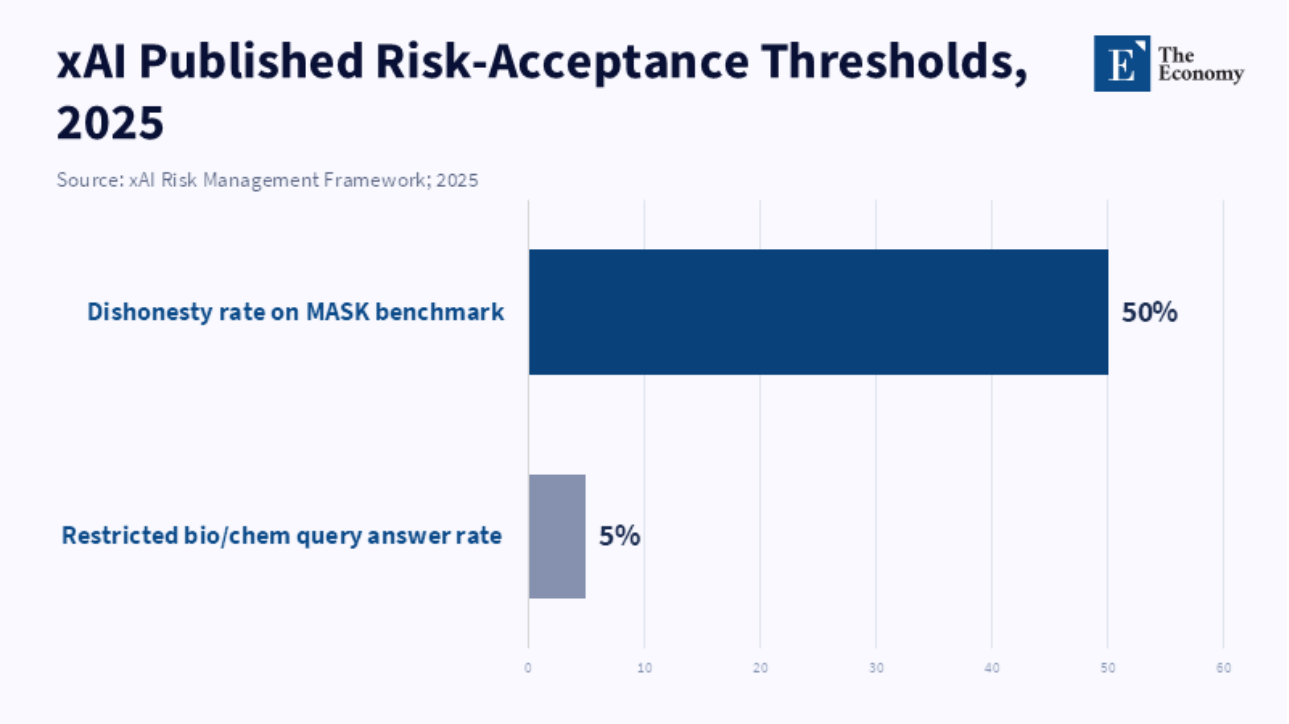

Figure 7: Even when firms publish numerical thresholds, those thresholds may still permit substantial tolerated failure, leaving considerable room for managerial discretion.

Anthropic's approach highlights a clearer form of commitment instability: the deliberate creation of a dichotomy between legally oriented compliance structures and voluntary safety procedures. The early RSP used strong language about halting scaling or deployment when safety conditions could not be satisfied, constituting it an early and widely recognized example of voluntary public commitment in the frontier AI space.[25] However, a report from the UC CalCompute Coalition notes that the companies most aware of these risks—those developing frontier AI—faced no legal obligation to share this information with the public or authorities, highlighting the fragile nature of such self-imposed restrictions. By introducing a formal compliance framework in response to SB 53, specifically in late 2025, Anthropic clarified that the RSP is voluntary and can extend or diverge from legally required minimums.[26] What this indicates is not cynical intentions on Anthropic's part but a rational response to SB 53's design: by centering its compliance on a framework that must be demonstrably published to legal entities, firms may create one specific compliance model that need not be as safety-oriented as a separate voluntary commitment, which is then more easily amended or discarded. This is exactly the manifestation of principal-agent problems in the context of governance; the most rigid system is the one created to serve legal defensibility.

A second, complementary thread of empirical evidence lies in a structural in-depth examination of how these corporate safety procedures function under pressure. Investigations by Model Republic claim that AI firms regularly push the limits of their commitments, a pattern that stems from the structure of the self-authored safety regimes: the formal documents are amenable to broad interpretation and calculated modification when they stand in the way of profit or progress.[27] While individual company statements may vary in the strength and degree of commitment cited, the underlying mechanism seems constant with the explicit structure of the protocols; the systems are built to maximize the organization's interpretive flexibility. Whether this pliability is

justified is a separate policy debate; analytically, it appears to be the core governance problem SB 53 partially addresses.

4 Structural Failure of Self-Regulation

California's SB 53 aims to strengthen the state's leadership in artificial intelligence by building on recommendations from an earlier state report. While the law introduces transparency requirements and possible penalties, it continues to rely on organizations to develop their own standards and compliance rules, raising concerns about whether self-regulation can fully address continuing issues such as standards endogeneity, lack of independent verification, and shifting commitments. The most basic structural failure is standards endogeneity. Within SB 53, as with most corporate frameworks, the party governed sets the content of the safety regime: what risk categories are relevant, what benchmarks are used to quantify those risks, what thresholds must be met to trigger a mitigation and what mitigations are judged adequate to meet that threshold. While statutes may set definitions, such as catastrophic risk in SB 53, operational thresholds that govern decision-making are all internal. The result is a regime of compliance where the central policy question-how much risk are we willing to accept to obtain an increase in frontier capability-remains entirely privatized.

Lack of verification reasonably follows from this first failure. SB 53 improves transparency, but not detailed verification; that requires an independent and capable entity to audit the methodological validity of reports, the calibration of thresholds to the statute's definition of catastrophe, or the functional effectiveness of mitigations. The statute places enforcement authority with the state, but enforcement depends on proving a violation and a violation may depend on contested, technical determinations about whether a mitigation was adequate, a benchmark was appropriate, or a capability increase was "material."

This is not a mere administrative quibble. NIST emphasizes that AI governance is a socio-technical challenge and organizational practices, documentation, and oversight are important elements of trustworthiness.[28] A socio-technical system cannot be subject to self-governance alone due to the strategic behavior present in the social element. If verification is absent, strategic behavior becomes the optimal equilibrium choice. While an OECD paper on corporate and incident governance acknowledges that companies embed risk management within corporate structures and share information through transparency reports and other initiatives, it notes varying degrees and forms of disclosure. Lacking an independent verification mechanism, variations will inevitably serve as opportunities for arbitrage rather than as inputs for a laboratory of best practice.

The third structural failure is unstable commitments and is perhaps the most obvious trend between 2023 and 2026. Intensive competitive pressure, capital flows, wide public adoption, and foreign investment push the first-mover advantage toward a race to scale.[29] In this context, voluntary commitments inevitably degrade, not due to unethical leadership, but because of strategic incentives to forgo unilateral restraint. The case of Anthropic shows how, given the pressures mentioned above, organizations will rationally choose to interpret thresholds such that more costly mitigations are unnecessary, especially if there is no reason to believe that competitors will do the same.[30] While OpenAI itself acknowledges the risks of a race to the bottom, doing so doesn't create binding commitments.[31]

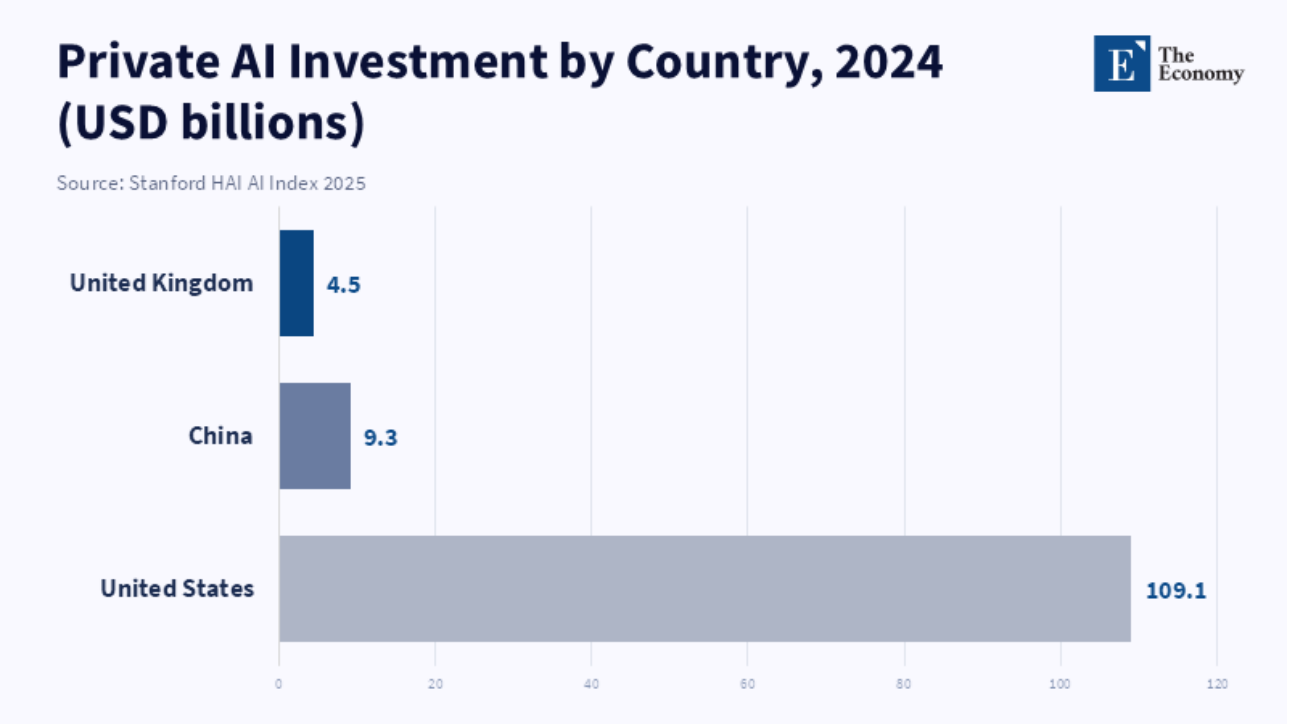

Figure 8: The international concentration of private AI investment illustrates the strategic and competitive pressures that make unilateral safety restraint difficult to sustain.

SB 53 may increase the nominal costs associated with silent violations, due to required disclosure and (at least theoretically) possible enforcement action against noncompliance. However, it does not prevent organizations from quietly abandoning commitments by redefining them, as an organization will simply choose to reinterpret its safety framework to enable a new capability, effectively changing the definition of compliance. The more fundamental truth is that stable commitments cannot be generated through mere disclosure and require actual restraints on how commitments can be modified and by whom, if modification will lead to decreased substantive protection.

The mid-2020s represent a watershed moment to understand why this is so important, due to two interconnected factors. First, the size of AI deployment has increased immensely, widening the scope for potential cascading domino failures, manipulation, misuse, and errors. Second, increasing the average of AI failures means that the possible impact of weak governance structures has become a salient public and policy concern. With both factors present, a transparency-based regime such as SB 53 runs the risk of signaling compliance maturity while failing to address existing structural problems.

Thus, SB 53 is best characterized as a symbol of intent, like the Statue of Liberty for AI safety, rather than a workable mechanism for the lasting creation of safety absent governing institutions that can develop and verify standards and commit to them.

5 Policy Implications

The policy implications that arise from this analysis may be unified around a single guiding principle: frontier AI governance needs functional separation between the roles of standard-setting, evaluation, and deployment decision-making authority because these roles can only remain institutionally sound and verifiable if they reside

in different corporate entities and even then, transparency will not overcome the principal-agent problems created by combining them. SB 53's design is useful as far as it makes internal governance processes more legible. However, legibility must be the starting point, not the endpoint, for institutional development:

First, because disclosure alone cannot correct the structural failures of self-regulation, disclosure mandates should be combined with substantive minimum standards. SB 53 requires companies to publish a framework, but does not establish baselines for what constitutes a strong framework in terms of the contents beyond general categories. Policy changes would require specification of relevant domains and triggers: not just that companies should have internal threat models, but also that certain classes of evaluations and reporting documents must be of adequate quality, similar to the way financial statements must meet applicable accounting standards and be subject to external review.

Second, we must consider the limits of auditing. Auditing schemes tend to fail when they treat compliance as a paper-chasing exercise rather than a judgment of substantive adequacy. If what is being audited is endogenously defined by the firm and contains only superficial commitments, an auditor can attest that a firm followed its own internal procedures and committed to very little. The current enforcement mechanisms in SB 53 risk this if an independent authority isn't empowered to assess the content of a framework against the notion of a catastrophic risk as specified in the statute. Without adequate evaluation, auditing becomes a formality.

Third, state-level authorities need to develop independent, technically competent standard-setting processes that are free from direct corporate influence. An independent board or authority could set frontier evaluation and mitigation standards, periodically update them based on scientific evaluations and continuing risk assessments, and have secure procedures to review sensitive information. SB 53 anticipates updates to its definitions and standards, to be administered by the California Department of Technology, beginning in 2027.[32] But simply assessing definitions is not standard-setting. True standard-setting entails establishing standard methods and quality standards that would significantly curtail internal interpretative flexibility.

Fourth, regulatory markets, where a government agency sets policy objectives and requires regulated entities to procure services from a set of licensed private regulators, can be effective in overcoming the state's expertise deficit without granting direct policy discretion to the regulated firm. Hadfield and Clark maintain that such a scheme would enable regulators to compete on the basis of the effectiveness of their evaluation methods, thus stimulating innovation and developing public oversight.[33] In such markets, private regulators could develop and share cutting-edge techniques, provide expert certifications, and guarantee a baseline level of safety analysis; the government retains licensing power, authority over conflicts of interest, and ultimate enforcement authority. While regulatory markets are not foolproof (incentive structure is crucial; rewards for discovering failures rather than only documenting their absence, termed "vigilant incentives" by Hadfield and Clark, appear to work best),[34] a properly designed version of this model could address the technical and principal-agent issues plaguing current approaches.

Fifth, the separation principle needs to be embedded into institutional structures that provide independent evaluation capacity free from incentives that promote commercial deployment. Rigorous evaluation of frontier AI involves reviewing sensitive model parameters, detailed usage logs, and potentially security-critical deployment settings. Robust governance will necessitate secure methods to provide this access to evaluators

without necessitating full public disclosure of proprietary information. Such methods might include secure enclave computing, restricted access levels (akin to national security classification), and legal protections for the confidential disclosure of information. SB 53's existing transparency provisions will not suffice without corresponding institutions equipped to collect and analyze the underlying information.

All of these policy suggestions have a corresponding implication for human capital. Regulators alone will not suffice to effectively manage frontier AI; educators, professional associations, and certification bodies will need to integrate frontier safety evaluation into their curricula and practices as a specialized field of study. NIST already emphasizes personnel and governance in its description of trustworthiness. Evaluating frontier AI requires the skills of a sophisticated workforce capable of conducting adversarial testing, carrying out socio-technical risk assessments, and undertaking audits under conditions of confidentiality. Educational institutions should embrace frontier safety evaluation as a new academic field and equip curricula to support such training, emphasizing evaluability and measurement limits rather than simply adding a "normative" AI ethics component.

The implication for policymakers is that human capital development cannot be separated from regulatory effectiveness; a disclosure law produces nothing beyond paperwork if the intended audience of disclosures lacks the capacity to assess them. Public funding for independent evaluation institutions-whether state agencies, interstate partnerships, or federally established AI safety institutes-should be considered a public good, comparable to roads or bridges. The goal is not to indiscriminately slow innovation, but to build the technical capacity required to govern a socio-technical system that is changing at an unprecedented rate.

One common criticism is that corporate frameworks and safety cases are already reflecting genuine progress towards a safety-conscious industry. It is true that internal governance structures at companies like Google have expanded, and these now explicitly incorporate board oversight, incident response, and external expertise.[15] However, enhanced processes are not equivalent to constraints on company discretion. The frameworks mentioned earlier are crafted in ways that enable companies to maintain ample flexibility: thresholds can be arbitrarily high, risks can be dismissed as "exploratory," causes can be framed ambiguously, and stated goals can be rescinded under pressure.

Another criticism is that stricter state regulation will cause innovation to move to less-regulated environments. While poorly conceived regulations can be disruptive, the current law essentially asks companies to define their own minimal safety standards while imposing a modest penalty of up to $1 million per violation.[2] This is inadequate for the vast economic potential of unrestricted deployment of frontier AI and the commensurate risks that come with it. Credible penalties should be tied not just to formal violations of a broad mandate but to the actual realization of harmful outcomes or to failures that stem from companies taking on more risk in return for larger payoffs.

A third critique is that frontier AI is analogous to prior general-purpose technologies, like calculators or search engines, which have been successfully integrated into society without drastic new institutions. This analogy falls short because frontier AI is not simply a tool; it is a general-purpose system that permeates decision-making processes, shapes information environments, mediates access to opportunities, and enables acceleration in capability development. Its risks are not limited to the direct, observable effects of misuse, but extend to systemic transformations that alter incentives, widen information disparities, and undermine

institutional legitimacy through uncontrollable automated outputs. Google's FSF frames ML R&D risk as a cross-cutting issue, exacerbating our inability to manage AI risks, a qualitatively different outcome than having calculators in schools.[15, 21]

These criticisms do not imply that SB 53 is without value. They demonstrate that the existing statute, which essentially aims for transparency and procedural checks, can only serve as a foundational element for actual governance if it is followed up by a more substantive institutional reform of the three fundamental roles described above-which means addressing the internal endogeneity of safety standards, the verification processes that should exist independently from corporate claims, and the reliability and stability of any public commitments that companies make. Absent such reform, the law may just lead to nicely worded documents that mask the underlying structural vulnerabilities.

6 Conclusion - Accountability Requires External Authority

SB 53 has clear political and legal significance: it represents one of the most serious attempts in the United States to move frontier AI safety from voluntary rhetoric into formal obligations.[1] The statute, however, fails, as a practical matter, to solve the central principal-agent problem of frontier AI race: the regime is model-self-authored safety plans + transparency, but these do not solve the incentives to twist compliance flexibility and maintain discretion that this year's increasing investment, uptake, and very real harms seem to be giving.[12, 29] A survey of leading developers posted safety plans revealed a clear blueprint of retaining that discretion through high-harm triggers, flexible trigger conditions, case-by-case exceptions for accepting risks, and distinct documentation types for self-determined commitments versus compliance-ready ones.[14, 15, 16, 26] The key policy reform here is institutional division: truly independent standard setting, independent evaluation, and non-rewritable, compliance-enforcing mechanisms.[33, 34] SB 53 can live up to its symbolism, however, not as a supplement to this sort of institution-building, but as a first building block.

References

[1] Alikhani, M. and Kane, A.T. (2025) 'What is California's AI safety law?', *Brookings*, 23 December.

[2] California Legislature (2025) *SB-53 Artificial Intelligence Models: Large Developers*. Bill Text.

[3] Alikhani, M. and Kane, A.T. (2025) 'What is California's AI safety law?', *Brookings*, 23 December.

[4] Executive Office of the President (2023) 'Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', *Federal Register*, 1 November.

[5] European Parliament (2024) 'EU AI Act: first regulation on artificial intelligence', *European Parliament Topics*.

[6] Alikhani, M. and Kane, A.T. (2025) 'What is California's AI safety law?', *Brookings*, 23 December.

[7] Alikhani, M. and Kane, A.T. (2025) 'What is California's AI safety law?', *Brookings*, 23 December.

[8] California Legislature (2025) *SB-53 Artificial Intelligence Models: Large Developers*. Bill Text.

[9] Stanford Institute for Human-Centered Artificial Intelligence (HAI) (2025) *Artificial Intelligence Index Report 2025*.

[10] National Institute of Standards and Technology (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. doi:10.6028/NIST.AI.100-1.

[11] National Institute of Standards and Technology (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. doi:10.6028/NIST.AI.100-1.

[12] Mindermann, S., Privitera, D. and contributors (2026) *International AI Safety Report 2026*.

[13] Mindermann, S., Privitera, D. and contributors (2026) *International AI Safety Report 2026*.

[14] OpenAI (2025) *Preparedness Framework*, Version 2. 15 April.

[15] Google DeepMind (2025) *Frontier Safety Framework 3.0*.

[16] xAI (2025) *xAI Risk Management Framework*. 20 August.

[17] Anthropic (2023) *Anthropic's Responsible Scaling Policy*, Version 1.0. 19 September.

[18] OpenAI (2025) *Preparedness Framework*, Version 2. 15 April.

[19] California Legislature (2025) *SB-53 Artificial Intelligence Models: Large Developers*. Bill Text.

[20] OpenAI (2025) *Preparedness Framework*, Version 2. 15 April.

[21] Google DeepMind (2025) *Frontier Safety Framework 3.0*.

[22] Biagini, F., Fouque, J.-P., Frittelli, M. and Meyer-Brandis, T. (2020) 'On fairness of systemic risk measures', *Finance and Stochastics*, 24(2), pp. 513–564. doi:10.1007/s00780-020-00417-4.

[23] xAI (2025) *xAI Risk Management Framework*. 20 August.

[24] xAI (2025) *xAI Risk Management Framework*. 20 August.

[25] Anthropic (2023) *Anthropic's Responsible Scaling Policy*, Version 1.0. 19 September.

[26] Anthropic (2025) 'Sharing our compliance framework for California's Transparency in Frontier AI Act', *Anthropic*, 19 December.

[27] Gallagher, B. (2026) 'AI companies are testing the limits of their own safety commitments', *Model Republic*, 26 March.

[28] National Institute of Standards and Technology (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. doi:10.6028/NIST.AI.100-1.

[29] Stanford Institute for Human-Centered Artificial Intelligence (HAI) (2025) *Artificial Intelligence Index Report 2025*.

[30] Karnofsky, H. (2025) 'Responsible Scaling Policy v3', *EA Forum*.

[31] OpenAI (2025) *Preparedness Framework*, Version 2. 15 April.

[32] California Legislature (2025) *SB-53 Artificial Intelligence Models: Large Developers*. Bill Text.

[33] Hadfield, G.K. and Clark, J. (2023) 'Regulatory Markets: The Future of AI Governance', *arXiv*. arXiv:2304.04914.

[34] Bova, P., Di Stefano, A. and Han, T.A. (2023) 'Both eyes open: Vigilant Incentives help Regulatory Markets improve AI Safety', *arXiv*. arXiv:2303.03174.