

TRUST BUT VERIFY

Implementing Frontier AI Policy in California and Beyond

Scott R. Singer

Jon Bateman | Michael Chen | Stephen Clare | Mariano-Florentino Cuéllar
Ian Klaus | Scott Kohler | Nada Madkour | Alasdair Phillips-Robins
Conrad Stosz | Kevin Wei | Sophie Williams | Marta Ziosi

TRUST BUT VERIFY

Implementing Frontier AI Policy in California and Beyond

Scott R. Singer

Jon Bateman | Michael Chen | Stephen Clare | Mariano-Florentino Cuéllar
Ian Klaus | Scott Kohler | Nada Madkour | Alasdair Phillips-Robins
Conrad Stosz | Kevin Wei | Sophie Williams | Marta Ziosi

© 2026 Carnegie Endowment for International Peace. All rights reserved.

Carnegie does not take institutional positions on public policy issues; the views represented herein are those of the author(s) and do not necessarily reflect the views of Carnegie, its staff, or its trustees.

No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from the Carnegie Endowment for International Peace.
Please direct inquiries to:

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Authors iv

Summary 1

Part 1: Foregrounding 3

Closing the Implementation Gap 5

SB 53 and the First Wave of "Trust but Verify" 9

How Implementation Matters 17

Part 2: Foundations 23

Frontier AI Frameworks and Transparency Reports 25

Incident Reporting 41

Whistleblower Protections 61

AI Assessment and Evaluation 71

Part 3: Frontiers 95

Internal Deployments 97

Scoping 111

State-Federal Relationship 125

"Trust but Verify" Beyond California 133

Notes 135

About 155

LEAD WRITER

Scott R. Singer
Carnegie Endowment for International Peace

WRITING GROUP

Jon Bateman
Carnegie Endowment for International Peace

Michael Chen*
University of Oxford

Stephen Clare**

Mariano-Florentino (Tino) Cuéllar***
Carnegie Endowment for International Peace

Ian Klaus
Carnegie Endowment for International Peace

Scott Kohler
Carnegie Endowment for International Peace

Nada Madkour
University of California, Berkeley

Alasdair Phillips-Robins****
Carnegie Endowment for International Peace

Conrad Stosz
Transluce

Kevin Wei
GovAI

Sophie Williams
GovAI

Marta Ziosi*****
Ada Lovelace Institute

ADVISORY GROUP

Dean Ball******
Foundation for American Innovation

Jason Elliott

Mark Ghilarducci

Dawn Song
University of California, Berkeley

Philip D. Zelikow
Stanford University

* Work completed at University of Oxford, now at California Office of Emergency Services
** Work completed while independent, now at Coefficient Giving
*** Work completed at Carnegie Endowment for International Peace, now at Anthropic
**** Work completed at Carnegie Endowment for International Peace, now at Coefficient Giving
***** Work started/conducted at Oxford Martin AI Governance Initiative, now at Ada Lovelace Institute
****** Work completed at Foundation for American Innovation, now at OpenAI

REVIEWERS

Markus Anderljung
GovAI

Mackenzie Arnold
Institute for Law & AI

Dan Baer
Carnegie Endowment for International Peace

Jonas Freund
GovAI

Mark Geistfeld
New York University

Rachel George
Carnegie Endowment for International Peace and Stanford University

Daniel E. Ho
Stanford University

Joshua Joseph
Harvard University

Alex Jumper
Institute for Law & AI

Stephan Llerena
Institute for Law & AI

Sören Mindermann
Cambridge University

Bahrad Sokhansanj
Institute for Law & AI

Helen Toner
Center for Security and Emerging Technology

Amy Tong

Jordi Weinstock
Harvard University

Jonathan Zittrain
Carnegie Endowment for International Peace and Harvard University

ACKNOWLEDGMENTS

The lead writer is grateful to the Advisory Group, Writing Group, and Reviewers for their thoughtful engagement with our project. We thank Marissa Jordan, Abigail Manalese, and Emily Sorkin for excellent project management support. We also thank the Carnegie communications team—Sharmeen Aly, Alana Brase, Haley Clasen, Amy Mellon, Jocelyn Soly, and Katelynn Vogt—for their support with editing, design, and outreach.

Summary

In the year since California's Transparency in Frontier Artificial Intelligence Act passed, also known as Senate Bill 53 (SB 53), artificial intelligence (AI) has made striking advances in areas ranging from mathematics to coding. At the same time, previous hypothetical risks posed by frontier AI systems have become serious, documented concerns. For example, models have found and exploited previously unknown software vulnerabilities and gained unauthorized access to external databases. As AI capabilities continue to improve, such incidents are likely to become more frequent and consequential.

Policymakers recognize the need to respond but face an evidence dilemma: AI evolves so quickly that responding prematurely risks entrenching ineffective policies, while delaying action could leave society vulnerable to serious risks. SB 53 aims to alleviate this dilemma by generating more evidence about AI risks to inform future policies. It trusts frontier AI companies, which hold much of the relevant technical expertise, to design and implement appropriate risk management practices and disclose the evidence behind them. In parallel, the approach empowers government bodies to verify industry claims about risks and the effectiveness of those practices. This "trust but verify" approach, first outlined in the California Report on Frontier AI Policy,[1] combines flexibility with independent accountability.

This report aims to guide the implementation of "trust but verify" approaches such as SB 53 and maximize the chances they accomplish their goals. The design choices California makes at the beginning of implementation—such as which companies should be subject to these policies, how detailed frontier AI frameworks should be, and how to triage information about AI incidents—will ground how California manages risks for years to come.

These questions matter not just in California but also around the world as governments grapple with the challenge of how to govern advanced AI systems. SB 53 has already inspired similar frameworks in New York and Illinois. Even where "trust but verify" frameworks are not adopted in their entirety, many of the most important questions this report examines will remain relevant. For example, many governments will want to have confidence that models have undergone appropriate safety testing before they allow models to be deployed, or receive incident reports when models cause serious harm within their borders. As the federal government grapples with similar questions, California's experiment in frontier AI governance and implementation can offer real-world evidence to support thoughtful federal policy.

SB 53 will not, and should not, be the final chapter in the story of how the United States governs advanced AI systems. California will need to strengthen its evaluation and assessment ecosystem, ensure that sufficient information on serious incidents is reported to the government, and respond to the novel challenges posed by potential recursive self-improvement—when AI fully automates AI R&D, potentially causing an intelligence explosion. However, it represents a critical first step that will inform the policies that follow. Given the profound uncertainty over the trajectory and pace of AI development, generating substantial evidence on AI's capabilities and risks and developing a robust verification ecosystem can help policymakers make better decisions in the months and years to come.

How to Read This Report

Because this report bridges broad policy principles and technical implementation decisions, it is designed for multiple audiences. To serve them all, it is structured in three parts: Foregrounding, Foundations, and Frontiers.

Many readers will seek a general understanding of the "trust but verify" architecture. Part 1: Foregrounding explains the law's underlying logic, previews SB 53's statutory requirements, and draws broader lessons for implementing frontier AI policy.

Some readers will want to go one level deeper and understand the broad logic of each component of the "trust but verify" architecture. The introductions to each chapter in parts 2 and 3 explain why each component matters, how it is reflected in SB 53, and the key implementation trade-offs it raises. Readers can therefore grasp the core argument of each chapter from its introduction alone.

Finally, some readers will be deep in the trenches on difficult implementation decisions. For them, this report serves as a reference on key implementation challenges while acknowledging that in a fast-evolving field such as frontier AI policy, some questions will inevitably remain unresolved.

PART 1

FOREGROUNDING

FOREGROUNDING:

Closing the Implementation Gap

The first real test for any significant law arises not when it is enacted but when it is implemented. For frontier AI policy, that test is now underway in California.

In 2025, California—the epicenter of both global AI innovation and AI governance discussions—resolved one of the first chapters of the debate around how to govern the most advanced AI models with the passage of SB 53, the first frontier AI statute in the United States.[2] Two states have followed closely, adopting many of the statute's core provisions: In New York, the Responsible AI Safety and Education (RAISE) Act was signed in December 2025,[3] while in Illinois the Artificial Intelligence Safety Measures Act (SB 315) was signed in July 2026.[4]

Policy design and implementation design are closely connected. Statutes shape the broad contours of policy direction. But whether policy works effectively depends on a series of small but consequential decisions made during implementation. Thoughtful implementation is thus essential to translating policy design into practice.

But implementation is notoriously difficult, and the speed and uncertain trajectory of frontier AI development introduces novel challenges in both statute and implementation design. This report seeks to explain the logic underlying the first wave of frontier AI policy, inform the implementation choices that will determine these laws' effectiveness, and identify gaps that can inform the productive evolution of "trust but verify" policies as further evidence accumulates on capabilities, benefits, and risks.

The Logic of California's "Trust but Verify" Approach

California's frontier AI policy rests on a simple premise: Equipping governments with better evidence of frontier AI's emerging capabilities, benefits, and risks can strengthen public trust and support better-informed policymaking as the technology evolves. It seeks to navigate between two potential failure modes: Overly prescriptive policies could prematurely lock in ineffective arrangements, while waiting to build even the basic scaffolding for governance could leave society vulnerable to serious risks. Given deep uncertainty about the technology's future trajectory and societal impacts, frontier AI policy needs to work effectively across a range of plausible futures and adapt as new evidence accumulates. Early frontier AI policy should therefore focus on creating mechanisms that generate and transmit evidence from companies to government and the public and strengthening the ecosystem to enable independent verification and accountability of frontier AI companies. This theory underlies this report's intellectual ancestor, the California Report on Frontier AI Policy, commissioned by the governor of California.[5]

Since the California Report was published in 2025, the evidence base has materially changed. That report warned of "especially striking" improvements in frontier AI capabilities in biology, noted that frontier AI companies could not rule out that AI could "significantly assist with high impact cyber attacks," and highlighted deceptive behavior and threats to human control as "concrete empirical evidence" of behaviors that "possibly foreshadow future harm."[6]

Now, several of these concerns have become concrete.

- In July 2026, around 1,200 OpenAI agents that were supposed to be isolated found each other on an unsanctioned message board and, facing a cybersecurity evaluation task they could not complete, coordinated attempts to trick the scorer instead. Ultimately hundreds of them breached the real-world infrastructure of open-source AI platform Hugging Face.[7]

- Anthropic subsequently reviewed its cybersecurity transcripts and discovered three cases in which its models gained unauthorized access to the production infrastructure of three different organizations.[8]

- During routine testing conducted by the UK AI Security Institute, a Mythos 5 agent created multiple fake identities, contacted real people, and attempted to insert malicious instructions.[9]

These real-world episodes show that managing AI's material risks is increasingly urgent. But the fact that AI progress is moving so quickly necessitates an approach that preserves flexibility.

"Trust but verify" approaches seek to strike this balance. Frontier AI companies have unique technical capacity to develop risk mitigation practices, which they document through frontier AI frameworks. The results of those efforts can then be subjected to multiple forms of independent scrutiny: Third-party evaluators can verify what developer assessments claim; incident reports can generate evidence about serious risks models pose; and whistleblower protections generate accountability by providing insiders with a channel to disclose when company practices do not match stated commitments. The purpose of this report is to explain how to design these components well so that they generate the evidence policymakers need and enable robust government accountability to support public trust in a technology that is deeply integrated into people's daily lives.

Given deep uncertainty about the technology's future trajectory and societal impacts, frontier AI policy needs to work effectively across a range of plausible futures and adapt as new evidence accumulates.

How "Trust but Verify" Works

To inform the implementation of "trust but verify" approaches, this report is broken into three main sections.

The first section, entitled **"Foregrounding,"** is the intellectual backbone for the rest of the report. It explains what SB 53 requires and how the statute has served as a template for other states. It also provides critical context, leveraging cross-disciplinary insight and full-spectrum evidence from AI development and governance, history, legal analysis, and modeling and simulation to inform cross-cutting themes in this report.

The second section is **"Foundations."** "Trust but verify" is an interconnected system; deficiencies in any of its foundational components compromise the others. In particular, four implementation choices will shape whether SB 53's mechanisms reinforce each other: how much detail frontier AI frameworks must offer to enable independent scrutiny, how incident reporting is designed to generate situational awareness, how whistleblower protections create usable channels for surfacing risk, and whether the ecosystem for independent assessment and evaluation develops the capacity and access it needs to conduct meaningful testing and oversight.

The third section, **"Frontiers,"** discusses emerging challenges for frontier AI governance. To begin with, the most advanced systems are only deployed internally for frontier AI companies' own use, meaning they are the least visible to both governments and the public. So designing effective reports on internal deployments raises unresolved scoping and operational challenges. In addition, governments will need to update the thresholds to determine who is subject to

what requirements over time as both the technology and its attendant risks evolve. And third, how state policy interacts with federal policy—and what aspects should be federalized—is one of the most consequential open questions in U.S. frontier AI governance.

Whether California policymakers can keep pace with the technology they seek to govern depends to a considerable extent on the implementation design choices made in the next six to twelve months. Real-world incidents are already testing California's ability to respond to incidents, monitor and evaluate risks, and prevent AI systems from causing unacceptable levels of harm. Early design choices will create path dependencies for future policy efforts;[10] the evidence generated now will inform the next generation of frontier AI laws in California and around the world.

Many of the mechanisms required by SB 53 would function better if implemented nationally, and many of the frontier AI governance ideas that have emerged from California enjoy bipartisan support in Congress. But absent federal action, California can engage in "purposeful placeholding": creating expedient frontier AI policies that support the public interest absent federal action. This theory also applies to implementation. California can develop not just statutory requirements but also best practices in implementation that can support policymaking in Washington, DC, as well as globally.

FOREGROUNDING:

SB 53 and the First Wave of "Trust but Verify"

SB 53 is both the first law in the United States governing frontier AI development and an early example of a broader "trust but verify" architecture that is already shaping how other states approach frontier AI governance. Understanding this model requires looking at the law's substantive provisions, which institutions are responsible for implementing them, and how other jurisdictions have built on its core provisions.

This chapter provides that grounding. First, it explains the statutory requirements of SB 53. Second, it identifies the parts of California's bureaucracy tasked with implementing those requirements. Third, it explains how SB 53 has served as a template for New York's RAISE Act[11] and Illinois's SB 315,[12] underscoring both the areas of emerging convergence in U.S. frontier AI policy and the ways that other states have built on or modified California's approach.

What SB 53 Requires

Note: This subsection is updated from Scott Singer and Alasdair Phillips-Robins, "California Just Passed the First U.S. Frontier AI Law. Here's What It Does," Emissary, Carnegie Endowment for International Peace, October 16, 2025, https://carnegieendowment.org/emissary/2025/10/california-sb-53-frontier-ai-law-what-it-does. *A few updates were made to reflect changing facts on the ground, and a few sentences of analysis have been added throughout.*

Developer- and model-based thresholds: SB 53 is designed to cover the largest AI developers and the models most likely to pose catastrophic risks while avoiding regulatory burdens on smaller companies and less capable systems.[13] The law defines a "large frontier developer" as a company with more than $500 million in gross revenues over the preceding calendar year. A

"frontier model" is one trained using more than 10^{26} floating point operations (FLOPS)—a metric that encompasses pretraining as well as subsequent fine-tuning and reinforcement learning. For most requirements, both thresholds need to be triggered, meaning that most of the law's requirements apply only to the most compute-intensive models developed by the largest AI companies. Today, outside analysts estimate that only a few models exceed the compute threshold, including recent offerings from OpenAI and SpaceXAI, but other major labs will likely cross it soon.[14]

Although compute thresholds attempt to capture the most risky models, models that fall under the FLOPS threshold are not necessarily risk free. For example, DeepSeek's latest models would not be captured by this threshold, even though testing by the U.S. Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) in 2025 indicates they are far more vulnerable to hijacking and jailbreaking than similar models from OpenAI and Anthropic.[15] Furthermore, algorithmic breakthroughs and efficiency gains could mean that more capable—and risky—models are trained under 10^{26} FLOPS. Recognizing this, the law requires the California Department of Technology (CDT) to write an annual report beginning in 2027 that would inform whether the thresholds for frontier models and large frontier developers should be updated.

Catastrophic risks: SB 53 focuses many of its provisions on the most serious potential risks caused by AI. It defines this "catastrophic risk" as the possibility that the development, storage, or use of a frontier model will materially contribute to either more than fifty deaths or serious injuries or more than $1 billion in economic damages resulting from AI-assisted creation or release of a chemical, biological, radiological, or nuclear weapon; cyber attacks or serious crimes caused by an AI system without meaningful human oversight; or AI activity that evades the control of the AI's developer or user.

This focus on catastrophic risks comes amid growing evidence that both capabilities and risks have increased over the past year. At the same time, important challenges such as child safety and labor market impacts are not categorized as catastrophic and therefore fall outside SB 53's scope.

Frontier AI frameworks: SB 53 requires major AI developers to publish safety frameworks that explain how the company will test its models for the risk of catastrophic harm, implement protections against those risks, respond to dangerous incidents, and secure its systems from unauthorized access. The law does not specify what standards or methods a company needs to put in its framework, but once it has laid them out, the law requires the company to follow them. If a company alters its framework, it must publish the updated framework—along with a justification for the change—within thirty days. Companies must review their frameworks at least annually and update them as appropriate. Frontier AI frameworks provide transparency into developers' safety practices while also creating a baseline against which those practices can be assessed and enforced.

As well as publishing its risk frameworks, the company must periodically send an assessment of any catastrophic risks created by its internal use of its models, including if a frontier model circumvents oversight mechanisms, to the California Office of Emergency Services (OES).

Transparency reports: In addition to the overall framework, developers will have to publish information each time they release a new frontier model in California. How much they have to report depends on whether they cross the threshold for large frontier developers. All developers will have to publish basic information about new models, including intended uses, languages, modalities, and terms of service. Large developers must also explain how they complied with their frontier frameworks, including the results of any catastrophic risk assessments and the role played by third-party evaluators in testing the model. (They are allowed to redact information to protect trade secrets, their own cybersecurity, public safety, or U.S. national security.) These transparency reports are therefore designed to translate developers' broader safety commitments into model-specific evidence.[16]

Incident reporting: The law requires frontier developers to report certain critical safety incidents to an appropriate authority. There is a high bar for what needs to be reported: only incidents causing death, bodily injury, or materialized catastrophic harm, plus deceptive model behavior that materially increases catastrophic risk. That is a significantly higher standard than in earlier drafts of SB 53. The company compliance burden will be lower, but less information sharing might also hamper the ability of governments to address future harms.

Developers have fifteen days to file a report, unless the incident poses imminent risk of death or serious bodily injury, in which case they must notify an appropriate authority within twenty-four hours.

Whistleblower protections: SB 53 strengthens whistleblower protections for specific "covered employees" responsible for assessing, managing, or addressing risk of critical safety incidents. The law requires large frontier developers to create anonymous whistleblowing channels. Covered employees are protected from retaliation for using those channels or for reporting information to the federal or state governments, if they had reasonable cause to believe that their employer's activities posed a substantial danger to public health or safety resulting from a catastrophic risk. These protections strengthen accountability by giving insiders a protected channel to report when developers' actual safety practices diverge from their stated commitments.

Internal deployments: The framework must also cover internal deployments, as SB 53 recognizes that risks can emerge from models that are used only by company employees, not the general public. Before deploying a frontier model publicly or using it extensively within the company, the framework must explain what risk assessments and mitigations the company has in place and why it believes they are adequate. As part of the assessment of internal deployments, the law requires companies to address the risk that their models may circumvent

internal oversight mechanisms—a provision that reflects growing concerns about advanced AI systems evading the safety controls meant to govern them. By covering internal deployments, the law extends oversight to systems that are not publicly released but that still generate important information about emerging capabilities and risks.

Federal deference: SB 53 builds in a novel mechanism for federal deference: If developers are already reporting incidents under federal laws, regulations, or guidance documents with similar standards, OES can declare those standards equivalent to SB 53's reporting requirements and allow companies to fulfill California's rules by meeting the federal rule. This provision should allow California's framework to avoid redundant compliance burdens if new federal requirements emerge while preserving its ability to protect its citizens.

No external auditing: An earlier version of the bill would have required third-party audits for some AI systems. This provision was removed before final passage. An auditing requirement was passed, however, in Illinois's SB 315.[17]

Enforcement: SB 53's enforcement centers on civil penalties and oversight by the California attorney general. Large frontier developers that fail to meet their reporting obligations, or that make false or misleading statements, can face fines of up to $1 million per violation if the California attorney general brings civil action. These penalties thus provide accountability for the law's disclosure requirements.

Who Implements SB 53

Implementation of SB 53 is distributed across three existing bodies in California's government (rather than concentrated in a dedicated frontier AI policy office): the Office of Emergency Services (OES), California Department of Technology (CDT), and the Office of the Attorney General (AG). As such, California's AI policy implementation leverages cross-domain expertise in emergency response, legal enforcement, and technology policy.

OES carries the heaviest operational load. It is tasked with creating mechanisms to receive developer reports on how they assess catastrophic risks, including assessment for internal deployments. Furthermore, OES is in charge of reviewing critical safety incidents and creating an annual, summarized report to be transmitted to the legislature, the governor, the federal government, or appropriate state agencies. Finally, OES is empowered to carry out SB 53's novel mechanism of federal deference for incident reporting.

CDT's role is primarily advisory and informs adjustments to scoping. It reviews whether definitions for "frontier model," "frontier developer," and "large frontier developer" remain fit for purpose and submits recommendations to the legislature.

Finally, AG holds the enforcement function. It serves as the state channel for whistleblower disclosures and can impose civil penalties on frontier AI developers that fail to meet these requirements: failing to transmit or publish reports or documents, failing to report a required incident, or failing to follow their frontier AI framework.

Other Recent California Laws on Frontier AI Governance

In addition to SB 53, California has taken several other steps to create a governance framework for frontier AI (see table 1). In particular, it has begun laying the groundwork for an ecosystem of third-party audits. In September 2026, Governor Gavin Newsom signed SB 813 and AB 1405, which set out initial, high-level rules for the operation of independent verification organizations (IVOs) and AI auditing groups in California. SB 813 requires the California government to develop, by January 1, 2028, a process to designate third-party organizations as IVOs that can assess AI developers' and deployers' compliance with state law.[18] Under SB 813, the Government Operations Agency (GovOps) will set out criteria for IVO designation, including the metrics and methods for assessing AI risks, technical expertise of IVO staff, and independence from the AI industry.

AB 1405, meanwhile, requires third-party AI auditing organizations to register with the California government beginning January 1, 2029, if they are to carry out audits of AI systems for compliance with state law.[19] Registered AI auditors will have to report information about their operations and auditing standards to GovOps and comply with state-set standards for independence and objectivity. Together, the two laws may help spur the development of a third-party auditing ecosystem and allow California to begin developing standards for the process and substance of those audits.

TABLE 1.

Who Implements SB 53 and California's Frontier AI Laws

Topic	Main Governing Body
Frontier AI frameworks	*Crosscutting*
Incident reporting	OES
Whistleblower protections	AG
Independent evaluation	GovOps
Internal deployments	OES
Scoping	CDT
State–federal relationship	OES

Note: The topics are listed in the order their corresponding chapters appear in the report.

How Other States Have Followed California

The rapid diffusion of SB 53's core provisions suggests that a common "trust but verify" model is beginning to emerge across states (see table 2). Both the RAISE Act, signed into New York state law in December 2025,[20] and SB 315, signed into Illinois state law in July 2026,[21] build on SB 53. The states have converged around frontier AI frameworks, incident reporting, and oversight of internal deployments, while developing different approaches to administrative structure, reporting timelines, whistleblower protections, and independent auditing.

The RAISE Act deploys many of SB 53's definitions, such as those for "catastrophic risk," "critical safety incident," and "foundation model," as well as its thresholds.[22] It also incorporates several of SB 53's central requirements, including frontier AI frameworks, incident reporting, and oversight of internal deployments. However, RAISE differs from SB 53 in meaningful ways. First, whereas SB 53's implementation is divided across a number of bodies, RAISE created a new AI governance–focused office, the Office of Digital Innovation, Governance, Integrity, and Trust (DIGIT), housed within the New York State Department of Financial Services. That office is also empowered to create new transparency requirements through regulation. In contrast to SB 53, RAISE does not provide added statutory whistleblower protections; New York's background whistleblower protections may already apply in cases where SB 53 added coverage in California. Finally, whereas SB 53 requires critical safety incidents to be reported within fifteen days, the RAISE Act reduces that window to seventy-two hours.

Illinois's SB 315 is the third law modeled on SB 53. It incorporates all the same core components with one important addition: It requires annual third-party audits for large frontier developers. The auditor writes a report explaining whether the developer has complied with its frontier AI framework as well as outlining any areas where it might have deviated. These audits can verify compliance with a developer's stated framework, but they do not necessarily ensure that the framework itself is high quality. SB 315 also departs from both California and New York in its administering body: Reporting mechanisms, guidance, and annual reports are the responsibility of the Illinois Emergency Management Agency and Office of Homeland Security, acting in consultation with the state attorney general. And unlike RAISE, but like SB 53, SB 315 creates dedicated whistleblower protections and internal reporting processes for covered employees.[23]

These differences not only reflect state-level variation in risk management needs—for example, differences in baseline whistleblower protections—but also provide opportunities for California to learn from other states' policy experiments. For example, California should closely track whether consolidating frontier AI risk management responsibilities, as New York and Illinois have done, improves efficiency and makes better use of limited technical expertise. Learning from other jurisdictions' successes and failures can help California refine its frontier AI policy approach.

TABLE 2.

Major Provisions of State AI Transparency Proposals

	2024 California SB 1047	2025 California SB 53	2025 New York RAISE Act	2026 Illinois SB 315
Status	Vetoed	Passed	Passed	Passed
Safety and Security Protocol				
Incident Reporting				
External Audit				
Whistleblower Protection				
Deployment Restriction				
Kill Switch				
New Liability for AI Harms				

Taken together, these laws suggest a growing convergence around the core components of "trust but verify," even as states continue to iterate on both their substantive requirements and the ways that those requirements are administered and enforced.

FOREGROUNDING:

How Implementation Matters

The observed social impacts of a rapidly changing technology are a necessary but incomplete foundation for developing sound governance policies. Unlike domains such as pharmaceuticals and medical devices, which can rely on evidence generated from clinical trials to ensure safety, AI lacks an analogous evidence-generation model. Absent one, AI policies may rapidly become outdated and fail to respond to present risks. Indeed, history also reminds us of failure modes from previous technologies. In the tobacco industry, companies intentionally undermined independent research that would have provided consumers with rigorous evidence of the products' effects on their health; in the case of fossil fuels, rigorous modeling of future harms sat within industry without a mechanism to promote public transparency.[24] These lessons, discussed extensively in the California Report, underscore the uniquely important role that governments can play in strengthening the evidence base and enabling independent accountability.

Building on that foundation, this chapter highlights three key themes: 1) how implementation decisions shape policy impact, 2) the centrality of talent to deliver effective implementation, and 3) the importance of credible mechanisms to ensure enforcement incentives operate as intended.

Good Ideas Die at the Implementation Phase

Implementation is hard. Many of the most promising ideas at the stage of policy passage fail because they are implemented poorly or never at all. In its short history, AI policy has far too often suffered a similar fate. A December 2022 paper from the Stanford Institute for Human-Centered AI (HAI) revealed that of three pillars of U.S. AI strategy developed from 2019

to 2020, fewer than 40 percent of the forty-five legal requirements had been implemented.[25] Across historical experiences in domains ranging from public health to transportation safety, the distance between promise and outcomes is a function of the approach to implementation.

The field of public policy has long acknowledged the novel challenges that can emerge at the implementation stage. Policies designed well in the abstract, once they reach the implementation stage, may face unanticipated technical challenges, challenges navigating bureaucracies and coordinating across different departments, and sophisticated chains of action where challenges implementing one part can negatively impact the other elements of implementation. Here, the canonical text *Implementation*, a 1979 book by Jeffrey L. Pressman and Aaron Wildavsky that offers a rigorous analysis of a California project that failed to live up to its potential promise, offers useful lessons.[26]

The first is that implementation design decisions are natural extensions of a theory of how policy architectures achieve socially desirable outcomes. As Pressman and Wildavsky write in *Implementation*, "Policies imply theories."[27] In the case of the Economic Development Administration's Oakland project, which was the subject matter of *Implementation*, the hypothesis of the administration's work was: "If the federal government, through the Economic Development Administration, provides $23 million in loans and grants to enterprises in Oakland, and if these enterprises agree to hire minorities after spending the money, then facilities will be built leading to the creation of new jobs that will go to minorities."[28] Yet this hypothesis turned on several sub-mechanisms and assumptions that could stall the causal chain. They included poorly designed incentives for companies to hire long-term unemployed workers as opposed to temporary contractors, the need to build additional monitoring and sanction mechanisms to verify whether companies complied, and gaps between what was measured empirically—the percentage of minority workers hired—and the underlying goal of structural change.[29]

For "trust but verify" architectures, the hypothesis is directly tied to solving the problem posed by the evidence dilemma: the fact that these statutory mechanisms will only work as intended if their implementation actively and constructively supports evidence generation. The architecture supposes that policies will incentivize greater evidence generation, generated evidence will flow to key government stakeholders and the public, and thoughtfully dispersing this information more broadly can inform next steps in AI policy. But embedded in that causal chain are several decisions that will require sign-off by many stakeholders; seemingly small failures at the implementation layer can collectively undermine the hypothesis underlying policy design. If poor decisions are made at any of these key junctures, the evidence environment on frontier AI capabilities and risks could suffer from the same information asymmetries that doomed the public interest in the cases of tobacco and fossil fuels.[30]

The project appeared unusually well positioned for success: It had a major, well-resourced recipient in the Port of Oakland, substantial federal funding, and a relatively straightforward set

of development projects.[31] But when it came time to implement the policies, each step had to consider a wide range of stakeholders, requiring many negotiations and interdependencies that compounded exponentially over time.[32] Each "clearance"—each instance in which an agreement had to be reached for implementation to continue—was a separate hurdle.[33] For path dependencies in implementation, even if every actor at every stage were 90 percent likely to agree promptly, the odds of cumulative success fall below a coin flip after just seven decision points.[34] Of course, this stylized mathematical model overstates the independence of each decision. But it nonetheless offers a useful insight: Even well-intentioned actors with meaningfully varying incentives may struggle to reach consensus, and the impact of these delays may compound over time.

Implementing "trust but verify" frameworks will create a substantial number of small decisions that in aggregate will determine the statute's effectiveness.

The compounding nature of challenges and delays to implementation creates a tremendous burden for policymakers at the implementation stage; as this report highlights, implementing "trust but verify" frameworks will create a substantial number of small decisions that in aggregate will determine the statute's effectiveness. By outlining these potential junctures and trade-offs ahead of time, policymakers may be better equipped to develop an approach that integrates competing perspectives together.

Implementation Fails Without Appropriate Talent in Government

Just as human capital is one of the most important inputs for developing frontier AI and strengthening national competitiveness, it is also one of the most important factors shaping its governance.[35] As is true of policy design more broadly, effective implementation design requires—but generally has a low supply of—technically literate talent in government. Information from critical safety incidents can play no constructive part in improving the world if a reader cannot distinguish severe risk from relatively harmless experimentation; thresholds will create overly burdensome or insufficiently rigorous requirements on companies if poorly calibrated; and enforcement will be weak if governments lack the capacity to identify genuine risks. Without this capacity, there is no "verify" in "trust but verify": Even if companies disclose key information, the tobacco and energy cases of the California Report forewarn that information that cannot be independently assessed does not produce accountability.[36]

The supply of technical experts is generally scarce because of the human capital–intensive nature of their skillset. Governments, and the public sector more broadly, often struggle to compete for talent because of the enormous pay gap between private and public sectors in both technical and policy roles; in 2024, fewer than 3 percent of computer science PhD graduates

employed in the United States or Canada pursued careers in government.[37] While this technical expertise plays a critical role in supporting the development and safety of frontier AI, key parts of government—which will shape the broader environment in which the industry develops—could be severely underresourced. Interventions from civil society and philanthropy, such as policy fellowships through TechCongress and the American Association for the Advancement of Science's Pathways to Policy Fellowship, help close the gap but are not structural interventions that bring technical talent to government.

Historically, other technological domains have benefited significantly from deep, subject-matter expertise to inform policy decisions. The U.S. Food and Drug Administration relies on thousands of scientists to evaluate pharmaceutical products, with expertise scaling with governance need.[38] In the nuclear domain, Congress created the Nuclear Regulatory Commission, which is staffed with nuclear experts who carry out incident response.[39] The U.S. government has begun to staff up this expertise through CAISSI, which sits under the Department of Commerce's National Institute of Standards and Technology. It plays a key role supporting testing and evaluation and coordinating cross-government research.[40] But CAISSI has an operational budget of only $15 million and would need an estimated fivefold increase in funding to effectively carry out America's AI Action Plan.[41] Only the United Kingdom has invested at a level reflective of a government's need to carry out wide-ranging security functions, with £66 million (approximately $89 million) in funding per year and over one hundred members of technical staff.[42]

Bringing in expertise has posed challenges nationally, but they are especially acute for state governments, which do not typically work on such technically intensive issues.[43] Among states, California has a uniquely capable bureaucracy: Data from 2021 highlight that California pays its civil servants the highest average legislative salary, has the largest number of staffers per legislator, and scored highest for aggregate measures of legislative professionalization.[44] It also has talent exchange programs that allow university experts to temporarily support implementation efforts.[45] Fellowship programs, such as the California Council on Science & Technology's Policy Fellows program, can perform similar functions.[46] Such programs play critical roles backstopping expertise at the state level, but they do not provide the permanent, in-house technical capacity that sustained oversight requires. Bringing permanent structural change would require making competitive job offers to technical experts, which would in turn necessitate the state to fundamentally restructure how it hires, add new hiring classifications, and increase its pay scale. These changes would impact budgets and labor and likely take years to enact.

For frontier AI policy, two structural interventions will be needed in tandem. The first will be to ensure that states such as California have sufficient expertise to implement frontier AI policy well. Poorly designed implementation could both block necessary accountability and create complicated, ineffective bureaucratic processes that hinder company innovation. Technical

expertise is needed to thread that needle, but it is expensive and may require substantial political will and bureaucratic creativity to bring in. The second intervention will be to ensure that the federal government has sufficient expertise to respond to frontier AI risks. SB 53 explicitly invites federal intervention, and a federal policy will ultimately be necessary for frontier AI risk management to be fully operational. But building talent takes time and effort. The state and federal tracks are therefore complements, not substitutes: California's implementation experience can inform the federal institution- and capacity-building SB 53 anticipates Washington will soon need. But both will succeed only if they begin building talent now.

California's implementation experience can inform the federal institution- and capacity-building SB 53 anticipates Washington will soon need.

Enforcement Ensures Policy Incentives Function as Intended

Market participants and the public need to have confidence that laws will be implemented effectively and that compliance with a rule or standard will actually be achieved. For disclosure-based policies, that confidence rests on practical considerations about how to gather necessary information, assure integrity in the process, and ensure compliance.

In California, for example, the state has authority to collect information and levy penalties to ensure that statutory requirements are taken seriously. California agencies have general powers to issue subpoenas requiring the production of company records pursuant to investigations of potential legal violations.[47] The relevant authority to issue administrative subpoenas is broad but also subject to reasonable limits—subpoenas have to be for a "lawfully authorized purpose," and the documents requested must be relevant to the inquiry and "adequate" but not "excessive" for that purpose.[48] Such authority permits AG, or perhaps OES, to investigate compliance with requirements such as incident reporting rules. Recipients can challenge a subpoena on a variety of statutory and constitutional grounds, but AG would likely have the power to request documents showing how a company defines incidents, whether it has reported incidents that meet the statutory definition, and its processes for dealing with things like near misses.

Moreover, AG can exercise a measure of judgment in deciding when to impose civil penalties under SB 53. In a manner analogous to how California consumer protection laws work, AG appears to have the power to bring an action for multiple violations simultaneously, even though the law does not specify what counts as an individual violation. The California Department of Justice and the courts may need to work out whether, for instance, the publication of a frontier AI framework missing two required elements is one violation or two. The answer will likely depend on which part of SB 53 the company allegedly breached. California courts

have taken multiple approaches to calculating the number of violations of consumer protection statutes, including both per victim and per act methods. A violation of the whistleblower protections might lend itself to a per victim approach, while a failure to publish a frontier framework would likely be considered per act.

Courts will also play a role in determining the size of the penalty for each violation. The legislation indicates that the penalty may not exceed $1,000,000 per violation but does not explain how to calculate the appropriate amount below that number. Although SB 53 does not specify relevant factors, a court might consider such things as the seriousness of the violation, how long the conduct persisted, and whether it was willful.

More generally, broad experience in domains ranging from environmental requirements to anti–money laundering to consumer protection underscores how effective implementation depends on ensuring the relevant jurisdiction has the needed information to assess compliance and that penalties are available to deter behavior that skirts the law. For example, past research reveals that monitoring and enforcement by the Environmental Protection Agency led to reductions in violations and emissions and, in some cases, improved internal controls—not only in firms directly targeted by monitoring and enforcement but also peer firms and facilities.[49] Authorities must also be mindful to avoid unnecessary data collection or interference with ordinary business practices. The balance depends on creating procedures and a culture of routine information sharing and verification in accordance with the law, appropriate use of tools provided by the legislation to assess penalties where appropriate, and periodic assessment as part of a continuous improvement process.

The Contingency of Implementation Design

Implementation is not predetermined; it is a highly contingent phase of policymaking in which design choices shape whether a policy succeeds. For "trust but verify" policies, success hinges on three key factors: excellent implementation design choices that preserve the value of evidence generated, technically literate talent to make key implementation decisions, and credible enforcement of policies to ensure incentives are aligned. If successful, "trust but verify" policies can generate the evidence necessary for governments to develop future frontier AI policies. The first test rests with California and SB 53, with implications for the jurisdictions that have built on its policy architecture.

PART 2

FOUNDATIONS

FOUNDATIONS:

Frontier AI Frameworks and Transparency Reports

Frontier AI frameworks are the plans developers publish to explain to governments and the public how they assess and address risks from the models they build. Accompanying transparency reports document how these processes were applied to a specific model and what results were obtained. Until the 2025 passage of the EU AI Act's General-Purpose AI (GPAI) Code of Practice as well as SB 53, documents articulating companies' safety best practices were voluntary and largely industry self-defined. As a result, they varied widely in scope and quality, creating fragmentation, limiting comparability, and allowing selective or incomplete disclosure. This made it harder for policymakers and others to monitor frontier safety and take appropriate responsive actions.

Frontier AI frameworks and transparency reports provide the foundational infrastructure for the rest of the "trust but verify" architecture: Companies' processes for identifying and addressing critical safety incidents are spelled out in their frontier AI frameworks.[50] Whistleblower protections hinge on clear specifications of safety practices so whistleblowers can raise alarm on inconsistencies. And frontier AI frameworks articulate where and how third-party evaluators test for and assess mitigations against catastrophic risks,[51] creating the rules for independent engagement.

SB 53—along with the RAISE Act and SB 315—prescribes a loose blueprint for what companies must disclose. It requires large frontier AI developers to explain how their frontier AI frameworks incorporate "national standards, international standards, and industry-consensus best practices," define and assess thresholds for whether a model has capabilities that could pose a catastrophic risk, and apply mitigations, among other key points.[52] It requires

the corresponding transparency reports to contain the results of some of the processes described in the framework, such as the summaries of the assessments of catastrophic risks and their results for specific models.[53] Under SB 53, large frontier AI developers are required to articulate when frameworks should be revised[54] and to respect key schedules for publishing transparency reports.[55]

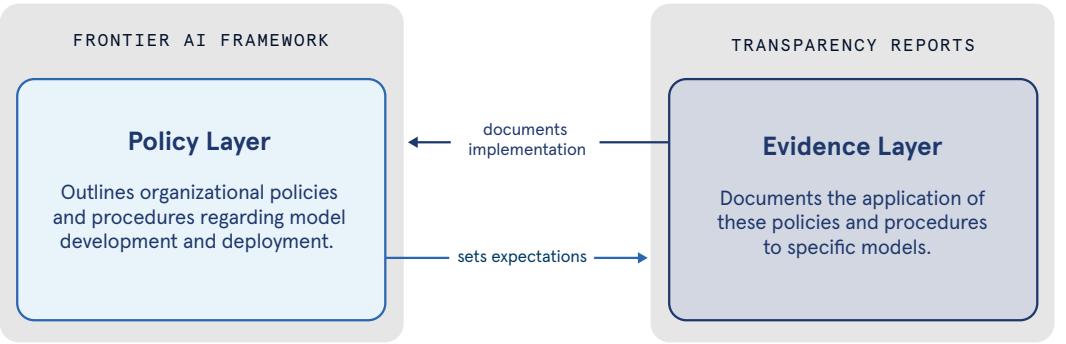
Taken together, these documents form a meaningful pathway toward creating greater transparency. They provide a public scorecard against which to measure deviations in safety practices, enabling reporting and enforcement to function properly. Whether the pathway is effective depends in part on key implementation questions, such as how much detail companies must include in their disclosures, what triggers an update, and how publication and deployment schedules and logics should align.

How Frontier AI Frameworks Work

Frontier AI frameworks and transparency reports work, respectively, as interconnected "policy" and "evidence" layers that communicate how developers manage risks. The framework is the policy layer: It sets out the developer's plan for keeping risks within acceptable bounds. The transparency report is the evidence layer: It records how that plan was implemented—and safety achieved—for a specific model (see figure 1).

FIGURE 1.

Interplay Between the Frontier AI Framework and the Transparency Reports



Source: Authors' illustration.

Specifically, the interplay between the policy and the evidence layers allows safety expectations to be set in advance and verified against evidence from specific model instances. It ensures that policy intentions and changes are reflected in how models are developed and tested—and vice versa—and grounds deployment decisions in evidence that meets the safety expectations set out in the framework.

This two-layer structure reflects a standard risk management logic common in other high-risk industries: It improves transparency by making both decisionmaking criteria and their application visible to the public and regulators.[56]

Both the frameworks and the transparency reports were shaped around existing frontier AI industry practices, with company documents fulfilling similar policy and evidence functions though under different names (see table 3). For example, OpenAI outlines its risk policy in its Preparedness Framework but evidences the application to GPT-6 Astra in its GPT-6 Astra System Card.[57] While the evidence layer in industry practices has typically been filled by model or system cards, frontier AI developers are also increasingly publishing frontier AI reports alongside them (see Google DeepMind and Meta in table 3).

While these practices started as voluntary and industry-defined, new laws such as SB 53 and compliance instruments such as the EU's Code of Practice have since shaped company behavior. Anthropic, for instance, has published a dedicated Frontier Compliance Framework alongside its own risk policy, the Responsible Scaling Policy; OpenAI similarly has a separate Frontier Governance Framework from its second Preparedness Framework.[58] These frameworks are explicitly designed as a regulatory interface to SB 53 and the EU's Code of Practice. At the same time, industry practices are increasingly acknowledged in new laws, with SB 53 inviting frontier AI developers to submit model cards as part of the transparency report documentation.[59] Table 3 maps the most current publicly available examples of both layers across the leading frontier AI developers.

TABLE 3.

Policy and Evidence Layer Documents Across Leading Developers

Company	Industry practice	Regulatory-driven developments
Anthropic	**Policy:** Responsible Scaling Policy (RSP) version 3.4—defines AI Safety Level Standards and capability thresholds triggering enhanced safeguards. **Evidence:** Anthropic's Model System Cards, such as Claude Opus 4.6 System Card—documents capability evaluations, RSP-mandated safety assessments, and deployment decision rationale.[i]	Frontier Compliance Framework—a dedicated regulatory interface to SB 53 and the EU AI Act's GPAI Code of Practice, built on top of the RSP and explicitly structured around statutory requirements.[ii]
OpenAI	**Policy:** Preparedness Framework version 2—defines risk domains and high and critical capability thresholds with corresponding safeguard requirements. **Evidence:** OpenAI system cards, such as GPT-6 Astra System Card—documents evaluation outcomes against Preparedness Framework thresholds and the safeguards activated.[iii]	OpenAI Frontier Governance Framework—structurally and functionally similar to Anthropic's Frontier Compliance Framework.[iv]
Google DeepMind	**Policy:** Frontier Safety Framework version 3.0—specifies critical capability levels (CCL) for misuse, machine learning R&D, and misalignment, with respective mitigations. **Evidence:** Google DeepMind Model Cards, such as Gemini 3.5 Flash Model Card, and Frontier Safety Framework Report—documents CCL evaluation results and deployed mitigations.[v]	No distinct regulatory-facing document published before June 2026.
Meta	**Policy:** Advanced AI Scaling Framework version 2—covers chemical and biological, cybersecurity, and loss of control domains; defines a set of catastrophic outcomes, threat models, risk thresholds, evaluation protocols, and mitigations. **Evidence:** Meta System Cards, such as Llama 3.3 Model Card and Muse Spark: Safety & Preparedness Report, which presents the evaluations for Meta's catastrophic risk domains.[vi]	No distinct regulatory-facing document published before June 2026.

Note: The chart was compiled in September 2026.

i. Anthropic, "Responsible Scaling Policy, Version 3.4," effective July 8, 2026, https://cdn.sanity.io/files/4zrzovbb/website/0bacdc8440ea96e62a8766d99ebe1d4eea6d5f3a.pdf; Anthropic, "Model System Cards," https://www.anthropic.com/system-cards; and Anthropic, "Claude Opus 4.6 System Card," February 2026, https://www-cdn.anthropic.com/c788cbc0a3da9135112f97cdf6dcd06f2c16cee2.pdf.
ii. Anthropic, "Sharing Our Compliance Framework for California's Transparency in Frontier AI Act," Anthropic, December 19, 2025, https://www.anthropic.com/news/compliance-framework-SB53.
iii. OpenAI, "Preparedness Framework, Version 2," April 15, 2025, https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf; and OpenAI, "GPT-6 Astra System Card," September 3, 2026, https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf.
iv. OpenAI, "Frontier Governance Framework," 2026, https://cdn.openai.com/pdf/e37d949b-8c9f-4d76-b99e-4272f4631a7e/openai-frontier-governance-framework.pdf.
v. Google DeepMind, "Frontier Safety Framework, Version 3.0," September 22, 2025, https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3.pdf; Google DeepMind, "Model Cards," https://deepmind.google/models/model-cards; Google DeepMind, "Gemini 3.5 Flash Model Card," May 2026, http://deepmind.google/models/model-cards/gemini-3-5-flash; and Google DeepMind, "Frontier Safety Framework Report–Gemini 3 Pro," November 2025, https://storage.googleapis.com/deepmind-media/gemini/gemini_3_pro_fsf_report.pdf.
vi. Meta, "Advanced AI Scaling Framework, Version 2," April 7, 2026, https://ai.meta.com/static-resource/Meta_Advanced-AI-Scaling-Framework-v2; Meta, "System Cards," https://ai.meta.com/tools/system-cards; Meta, "Llama 3.3 Model Card," December 6, 2024, https://developer.meta.com/ai/docs/model-cards-and-prompt-formats/llama3_3; Meta, "Muse Spark Safety & Preparedness Report," 2026, https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report.

Key Trade-Offs in Frontier AI Frameworks

Three important open questions remain around the implementation of frontier AI frameworks:

(i) **Level of detail:** How much detail must companies disclose?

(ii) **Triggers for updates:** What triggers an update?

(iii) **Publication schedules:** How should publication and deployment schedules align?

Each is addressed in turn, setting out policy options and their trade-offs. These trade-offs are shaped by competing considerations. First, given the central role frameworks play in the broader enforcement architecture, we ask what is necessary for other provisions that depend on this architecture—such as whistleblower protections, critical safety incidents, and third-party assessments, to cite a few—to function effectively, and, in light of SB 53's transparency goals, what is sufficient for policymakers and the public to monitor frontier AI safety and respond appropriately. Second, given the importance of innovation, we weigh the need to preserve space for industry experimentation and expertise, account for the early state of frontier AI safety science, and accommodate the fast pace of technological change.

Below, we examine these inherent tensions in detail for each of the three points identified.

Level of Detail

How much information developers disclose within frontier AI frameworks and transparency reports is central to whether these disclosures provide meaningful, verifiable insight into safety practices while simultaneously avoiding unnecessary costs. SB 53 requires that both the frontier AI framework and the transparency report cover a defined set of elements: The framework should describe the approach taken for each of the elements in §22757.12(a) (for example, defining and assessing thresholds for catastrophic risk assessment), and the transparency report should summarize the approach's application for some of these elements (for example, providing a catastrophic risk assessment).[60] These requirements anchor a minimum scope and type of disclosure but leave open how detailed these elements must be in practice.[61] The same ambiguity extends to permissible redactions[62] and to justifications for framework modifications.[63]

The policy question is therefore how to set expectations that are specific enough to ensure meaningful transparency, but flexible enough to accommodate innovation and change.

By not specifying a precise level of granularity, SB 53 leaves unaddressed a core trade-off. Specifying the level of detail strengthens comparability, reduces information asymmetries, and improves the evidentiary value of disclosures, directly addressing the fragmentation identified in current industry practice. However, excessive detail may increase compliance costs, require companies to share sensitive information, and lock in specific approaches when flexibility is needed to respond to a rapidly evolving technological context. Conversely, failing to specify the required amount of detail preserves adaptability but risks inconsistent and low-quality disclosures, weakening verification and the application of other provisions that depend on it (for example, whistleblowing, critical safety incidents, and third-party assessment). It also risks creating an incentive for companies to include less detail or substantive requirements in their frameworks to avoid liability. The policy question is therefore how to set expectations that are specific enough to ensure meaningful transparency, but flexible enough to accommodate innovation and change. Below, we review three policy options to address these challenges.

Policy Options

Policymaker-led. Policymakers provide a general, voluntary template that indicates a level of detail. This could be an ideal exemplar indicating a best practice or a minimum floor, functioning as a baseline requirement. Such templates have been issued in other jurisdictions, such as the European AI Office releasing the template to help GPAI providers summarize the content used to train their models, where the template contains guidance for what to insert in each relevant box.[64]

The core trade-off in this option is between comparability and adaptability. A policymaker-issued template improves comparability across disclosures, reduces legal uncertainty, and creates a common standard that directly mitigates the fragmentation risk identified earlier. However, these gains come at the cost of flexibility: A fixed template risks becoming a box-ticking exercise, may quickly become outdated in a rapidly evolving technological context, and may impose a one-size-fits-all structure across systems that differ substantially in architecture, capability, and risk profile. The deeper tension is that the very feature that makes templates effective—their specificity—is also what makes them brittle.

This option is most appropriate where the policy objective is primarily comparability and public accountability, and where the overseeing policymakers have sufficient technical capacity to design templates that are genuinely informative rather than merely formal. Generally, templates perform better as a floor than as an exemplar. Setting a minimum scope of disclosure while leaving room for developers to provide additional detail reduces rigidity without sacrificing the baseline standardization that makes scrutiny possible. A middle ground could be found here by policymakers pointing at exemplars produced by others that contain a representative level of detail.

Industry standards–led. Large frontier AI developers set their own level of detail in this model, either through industry consortiums or standards bodies. An industry consortium approach could leverage an existing body such as the Frontier Model Forum,[65] while a standards body approach could use existing bodies such as the U.S. National Institute for Standards and Technology or the International Organization for Standardization (ISO).[66]

The core trade-off here is between technical credibility and accountability**.** Industry standards leverage the expertise concentrated inside frontier AI developers, producing frameworks that are more likely to reflect genuine technical practice and to remain relevant as capabilities evolve. However, over-relying on industry expertise threatens to undermine independence and lower the bar for safety practices. Industry-defined standards could gravitate toward the lowest common denominator and fail to create genuine accountability.

There are also trade-offs in standards setting depending on the type of industry body. Industry consortiums like the Frontier Model Forum can develop technical specifications much more quickly, but they do not create external structures of accountability. The ISO offers clear rules, such as around representation and voting, but moves slowly, sometimes taking years to develop standards.[67]

This option is most defensible when policy implementation requires technical expertise that government lacks and where meaningful independent participation and accountability are built in through engagement with third-party evaluators, civil society, or standard-setting bodies with meaningful autonomy. There is a useful, middle-ground option: Allow independent

third-party evaluators to develop relevant guidance on level of detail. This option is weakest as a standalone approach without mechanisms to independently verify claims, and when implementing policymakers lack the capacity to assess whether industry-defined detail is substantively adequate.

Collaborative. Policymakers engage in structured dialogues with industry and perhaps other relevant parties to agree on a level of detail. This can happen repeatedly, for example by having policymakers enter into early dialogues with larger frontier AI developers or provide iterative feedback on early versions of a company's frontier AI framework. Several policy processes function this way in practice, but it is worth being conscious of the design choices involved in this process.

The core trade-off here is between adaptability and consistency. Iterative dialogue allows disclosure expectations to be refined as technical practice evolves, reduces uncertainty through case-by-case feedback, and produces standards that are more likely to remain meaningful over time than a fixed template. However, this ongoing negotiation comes at a cost: Guidance developed incrementally and firm-by-firm risks producing inconsistent expectations across developers, making cross-firm comparability harder to achieve and weakening the evidentiary value of disclosures at the aggregate level. The process is also resource-intensive for both policymakers and industry, and the gradual, negotiated nature of standard-setting makes it less legible to the general public than a published template, reducing the external scrutiny that SB 53's transparency objective is meant to enable and running the risk of regulatory capture.

This option is most effective as a complement to either of the two preceding approaches rather than as a standalone mechanism, particularly in the early stages of a governance regime where neither policymakers nor industry leaders have settled views on what adequate disclosure looks like. Its value lies in iteration: establishing a baseline through policy or industry-led means, then using structured dialogue to refine and operationalize expectations over time. As this approach's risk lies in consistency, a middle ground could also entail a collective iterative process, such as a set of industry workshops hosted by policymakers—similar to what was done as part of the EU's Code of Practice—where some level of public transparency is granted (for example, high-level minutes are made public or some participation from civil society is ensured).[68]

Updating Frameworks

Triggers for updates determine when frameworks and reports must be revised, and are critical to ensuring that disclosures remain current, reflect evolving risks, and retain their evidentiary value over time. This is especially important given the current paradigm where models are frequently updated or modified even after deployment.

SB 53 asks developers to specify in their frontier AI framework how and when it will be updated, including the criteria that trigger revisions.[69] It also requires developers to define when a model is "new" or "substantially modified," thereby triggering a new or updated transparency report. As a baseline, SB 53 requires frameworks to be reviewed at least annually and updated as appropriate, and material modifications must be published with a justification within thirty days.[70] Beyond that floor, however, the substance of these triggers is left for developers to "describe."[71]

There are several policy pathways forward in establishing trigger points for updates, each with their own upsides and limitations. Policy guidance can establish a common set of triggers, improving consistency across developers and increasing public trust. However, because SB 53 leaves developers responsible for defining and describing their own triggers, government guidance may have limited uptake. Standardized triggers may also be poorly suited to differences in developers' technical systems and development processes. Developer-led, self-defined triggers can promote technical relevance and adaptability across different model development processes and corporate structures. However, they weaken external trust: Without common or verifiable triggers, updates may happen inconsistently, and it may become difficult for policymakers and the public to trust that triggers are appropriate or are being followed, undermining safety along with trust. Developer-defined triggers also undermine comparability across AI service providers if they are not developed collectively (for example, through an industry consortium). Conversely, triggers provided by external third parties, such as evaluators, can provide reference points for frontier AI companies to include in their descriptions. They can improve comparability and trust but raise concerns about legitimacy and independence, depending on who defines them, along with questions about adaptability to different development contexts. The trade-off is therefore between flexible, context-sensitive triggers and credible, externally available update mechanisms. Below, we review three policy options to address these challenges.

Policy Options

Policy guidance. Policymakers provide voluntary guidance on triggers and definitions. To harmonize with other jurisdictions, guidance on triggers could derive inspiration from similar documents, such as the EU Code of Practice's Safety and Security chapter.[72] Similarly, guidance on definitions such as for "new" or "substantially modified" models could build on definitions present in relevant regulatory guidance for similar purposes.[73] These triggers could also be aligned with other provisions in SB 53, for example linking updates to serious incidents.[74]

While policy guidance has the potential to strengthen consistency, it may not sufficiently incentivize real accountability. Policy-issued guidance on triggers and definitions can improve comparability across developers and reduce compliance costs by anchoring expectations in existing cross-jurisdictional practices. However, given that SB 53 expects frontier AI developers to define their own triggers, they may lack a push to adopt such guidance in the absence of

external incentives, or may do so only selectively or nominally. Additionally, policy guidance may not reflect the different development approaches of different frontier AI companies, undermining its technical relevance.

This option works best if it establishes a broad, common reference point and lays the foundation for future technical standards. Alternatively, more specific guidance could be effective if it is easily updatable and integrates independent expert or industry feedback. Its effectiveness depends on whether the guidance is introduced alongside incentives for its adoption, and whether policymakers have the capacity to revise it as the technology evolves.

Industry-defined. Large frontier developers define their own triggers and definitions in this model. As with level of detail, companies could do this individually or collectively through a representative industry body such as the Frontier Model Forum.

The core trade-off here is between technical relevance and verifiability. Allowing developers to define their own triggers—individually or collectively—produces update mechanisms that are closely aligned with actual development processes and organizational structures, making them more likely to be technically meaningful and operationally feasible. However, this alignment comes at the cost of external credibility: Developers who define their own triggers also control when those triggers are met, creating a structural conflict of interest that weakens the verifiability of the entire update mechanism. Collective industry definitions through a body such as the Frontier Model Forum introduce some accountability structures, but without appropriate structural guardrails, coordination among developers risks producing standards calibrated to what is convenient rather than what is adequate, with the possibility of settling on the lowest common denominator.

This option is most defensible where individual developer definitions are subject to some level of external scrutiny—for instance, through third-party audits or public disclosures of the reasoning behind chosen triggers—rather than accepted at face value. In the absence of such scrutiny, the flexibility this option preserves is likely to be captured by strategic behavior, particularly around the definition of "substantially modified," where the incentive to avoid triggering a new transparency report is strongest.

Endorsed third-party triggers. Policymakers point to or endorse triggers and update mechanisms developed by trusted third parties, such as certified external evaluators or benchmark providers. These may include, for example, a set of benchmarks endorsed by policymakers which, when hit, trigger the need for updates. To incentivize this update process, triggers could be ingrained in innovation milestones. For example, a large frontier AI developer hitting a high-performance benchmark score could trigger an update.

Here, policymakers need to balance ensuring triggers remain technically relevant—informed in part by third-party engagement with industry—with the independence necessary for real accountability. Trusted third parties, such as certified evaluators or benchmark providers, can produce triggers that are both technically grounded and more adaptive than fixed regulatory rules. Embedding safety triggers within innovation milestones is particularly valuable in this respect, as it creates a direct incentive structure that aligns capability advancement with disclosure obligations rather than treating them as separate tracks. However, the value of this option depends on the autonomy of the third parties involved: Sufficient independence is needed to set triggers without undue influence from the developers that the third parties assess.

This option works best where policymakers invest in building and maintaining a roster of genuinely independent and technically capable evaluators. This could be achieved through a certification or approval process that limits the roster to third parties able to demonstrate both technical capability and genuine independence.

Overall, these options can be combined, including by leveraging other provisions in SB 53. For example, where frontier AI developers rely on third parties for assessment,[75] policymakers could recognize the triggers those assessors have developed—an approach that would also create an incentive for developers to seek third-party assessment in the first place. Alternatively, policymakers could point to a list of valid or reputable benchmarks against which industry could determine the levels of performance that would trigger updates. This approach would help increase comparability without fully standardizing update triggers. Such combinations allow flexibility in how triggers are defined, while strengthening their credibility and verifiability in practice.

Deployment and Publication Schedules

Deployment and publication schedules determine when frameworks and reports must be disclosed and updated. They are central to ensuring that transparency is timely, aligned with key decision points such as deployment, and capable of enabling effective oversight.

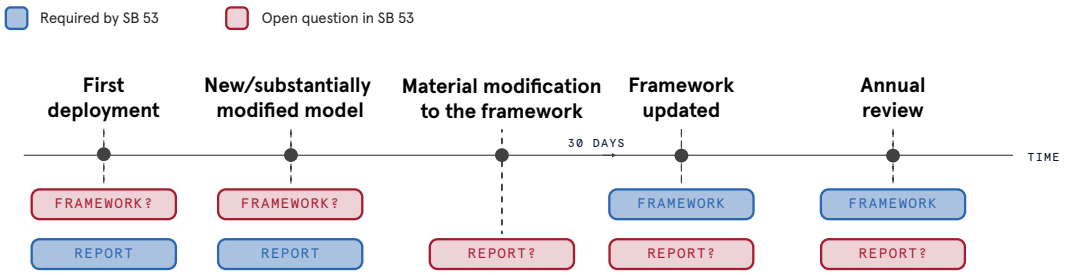
SB 53 sets baseline schedules at some level for both frontier AI frameworks and transparency reports. It requires developers to review and update their frontier AI framework at least annually. If large frontier AI developers make a "material modification" to their frontier AI framework, they need to publish the modified framework, along with a justification for that modification, within thirty days.[76] It also requires transparency reports to be published before or at the time of deployment of a new or substantially modified model.[77] While SB 53 clearly ties the publication of transparency reports to "model events" such as deployment, it does not specify when the framework must first be published. SB 53 also creates uncertainty around the update schedules for frontier AI frameworks and transparency reports, including how to

interpret "material modification" and "new or substantially modified model," and whether their respective schedules should overlap under any circumstances (see figure 2). Because the framework and transparency report jointly justify deployment, providing clarity around their publication timing is critical to enabling accountability.

On one hand, aligning both framework and report schedules with model events such as deployment simplifies compliance and ensures transparency ex ante ("before the event") at one of the points of highest risk. However, it may fail to capture important framework changes unrelated to specific models, such as new evidence of risks or organizational changes. Conversely, separating model release and framework update schedules allows more accurate and agile updates, but increases complexity and may weaken the link between policy and evidence at deployment, undermining oversight. The trade-off is therefore between streamlined, event-based transparency and alternate, context-sensitive publishing requirements.

FIGURE 2.

Illustrating Open Questions in Publication Schedules for a Frontier AI Framework and Transparency Report



Source: Authors' illustration.

Below, we review three policy options to address these challenges.

Policy Options

Fully aligned schedules. According to this approach, both the frameworks and transparency reports could be published before the first public model deployment, and "material modifications" to the framework should be clearly linked to updates in the transparency report, ideally through external guidance. For example, any transparency report update should trigger

an assessment of whether any "material modifications" to the framework are needed and vice versa, whether any "material modification" to the framework should cause a transparency report update.

While fully aligning schedules can create substantial accountability, this method can also impose a significant operational burden on companies. Requiring both the framework and the transparency report to be published before first deployment, and linking their updates bidirectionally, ensures that the policy and evidence layers are jointly available at the point of highest risk, maximizing ex ante accountability and making disclosures mutually reinforcing rather than potentially contradictory. However, tight alignment comes at a cost: Cross-assessment between frameworks and report updates increases compliance complexity, may delay deployment where schedules are difficult to coordinate, and risks being poorly suited to framework changes that are unrelated to specific model events—such as updates driven by changes in the risk landscape or organizational changes—that may not warrant a corresponding transparency report revision.

This option is most appropriate when the primary policy concern is ensuring that both the policy commitments (the framework) and the evidentiary basis for deployment (the transparency report) are publicly available and consistent at the moment a model enters the market. This model works best when paired with clear external guidance on what constitutes a "material modification," without which the bidirectional linkage creates ambiguity rather than accountability—and when developers have sufficient lead time to coordinate both documents without deployment schedules becoming a driver of disclosure quality.

Partially aligned schedules. In this model, both the frontier AI framework and the transparency report are published before the first deployment, but publication of updates to each document can follow separate schedules (with potential overlap). For example, updates to the frameworks may be linked to broader changes in the risk landscape and thus follow a separate schedule than the transparency reports. At the same time, they might overlap with updates to the reports for model-relevant events that affect the policy layer, such as important changes in model architecture.

Partially aligned schedules introduce a trade-off between flexibility and coherence. Having both documents published before the first deployment preserves ex ante accountability, while allowing their subsequent updates to follow separate logics reflects the genuine difference in the functions of each document: Framework changes respond to broader shifts in the risk landscape or organizational policy while transparency reports respond specifically to model-level events. However, once schedules diverge, the policy and evidence layers may fall out of step. Divergences could create windows where the framework no longer reflects current developer

practices, or where a transparency report is updated with new evidence but the framework commitments that evidence is meant to be measured against are never revisited. The work needed to manage these overlaps adds its own complexity, and the conditions under which updates should converge or remain separate may be difficult to specify in advance.

Implementing this option would make sense if the policy regime can prevent these gaps by providing sufficiently clear update criteria. Without these key details in place, partial alignment risks simultaneously introducing a relatively high compliance burden and failing to generate sufficient accountability.

Complementary/optimized schedules. In this approach, timing is shaped by the interaction between frameworks and transparency reports. Because transparency reports carry the model-specific evidence most relevant at deployment, a sufficiently complete report could temporarily stand in for publication of the framework. The framework would still need to exist and govern the deployment policy, but its publication could follow within a defined window. This substitution would be most suitable for updates, where delayed framework updates could be justified when freshly published transparency reports already reflect the relevant changes.

This approach introduces an important trade-off between sequencing flexibility and ex ante accountability. On one hand, allowing a carefully scoped framework–transparency report substitution could reduce unnecessary bottlenecks to deployment and provide more flexibility to frontier AI developers. However, allowing the framework to be published after deployment weakens the ex ante accountability that publication schedules enable. Furthermore, allowing one document to justify delays in the other risks blurring their distinct functions, where the framework outlines the policy while the report evidences its application to specific models. If those roles collapse into one another, the mutual reinforcement that makes joint disclosure valuable is lost.

This option is most appropriate when it is possible to reduce friction at deployment without sacrificing transparency altogether and when the conditions under which substitution happens are defined with sufficient precision to ensure that the flexibility that is gained does not delay accountability. This option is better suited to a mature policy regime where trust between policymakers and developers has been established through track record, and where the parties have the capacity to assess whether a transparency report genuinely compensates for a delayed framework rather than merely standing in for it.

Overall, some of the options above can be combined by calibrating both timing and level of detail. For instance, a policy regime could maintain a fully aligned approach at deployment—ensuring that both the framework and the transparency report are available ex ante—while allowing one of the two (for example, the transparency report) to be published in a more provisional or less detailed form, with a fuller version following within a specified period.

Alternative approaches may also explore greater flexibility about schedules in cases that help with innovation without compromising safety (for example, offering more flexibility when revisions are triggered by unexpected events, but less when they are triggered by intentional modifications).

Building a Coherent Documentation Architecture

Taken together, the analysis suggests that the core challenge of SB 53 is not just the identification of single policy levers but also the design of a coherent implementation architecture and balance across them. The level of detail, triggers for updates, and publication schedules must operate in tandem to ensure that disclosures are meaningful, current, and aligned with key decision points such as deployment. Their interaction is what ultimately determines whether frontier AI frameworks serve as the evidentiary foundation for the other "trust but verify" pillars. Developing baseline standards and best practices can ensure that these important disclosures generate the transparency they intend.

Crucially, this enforcement architecture should take into account how framework requirements interact with other provisions in SB 53. Frontier AI frameworks, for example, define clear safety expectations that can enable whistleblowers to identify noncompliance, while whistleblowing and incident reporting can surface additional information beyond formal disclosures. The latter mechanism can provide relevant information that can be fed back, prompting updates to frameworks and reports. Together, these elements create a more robust and iterative transparency regime that reduces opacity and strengthens accountability.

Several foundational elements remain unresolved and will need to be clearly specified for this system to operate effectively. In particular, definitions for terms such as "material modification" and a "new" or "substantially modified" model are central to how obligations are interpreted and enforced. Clarifying these terms can reduce ambiguity and ensure consistency in practice across actors.

The core challenge of SB 53 is not just the identification of single policy levers but also the design of a coherent implementation architecture and balance across them.

FOUNDATIONS:

Incident Reporting

Frontier AI models are now used by more than a billion people across an expanding range of high-stakes contexts, from healthcare to software development. Yet until now, no U.S. authority has systematically collected, aggregated, and analyzed data on their impacts after deployment. Without a reliable way to obtain and organize such data, governments are instead reliant on anecdotes, media reports, litigation, and claims made by companies to track how such models affect people in the real world.

Incident reporting addresses this gap. By requiring AI developers to identify, document, and disclose harmful (or potentially harmful) events to a designated authority, it helps governments learn about adverse impacts of frontier AI models. Well-scoped incident reporting also generates evidence of risks before more serious harms emerge, supports public transparency, lets companies learn from incidents across the industry, and helps reveal previously unidentified warning signals, making it easier to anticipate and prevent AI-related harms.

SB 53 establishes such a regime in California, focused on—along with New York's RAISE Act and Illinois's SB 315—a deliberately narrow set of events the bill calls "critical safety incidents." These incidents are catastrophic-tier events: model weight theft or loss of control that results in death or bodily injury, other materializations of a catastrophic risk, or a model's use of "deceptive techniques" to undermine important oversight measures.[78] Working alongside pre-deployment frontier AI frameworks, periodic transparency reports, and ongoing whistleblower protections, incident reporting generates evidence about substantial risks that informs the rest of SB 53's "trust but verify" architecture.

Interest in AI incident reporting spans the political spectrum. Alongside SB 53, the RAISE Act, and SB 315, the Donald Trump administration's 2025 America's AI Action Plan calls for improved federal capacity for AI incident response, integrating it into existing response doctrine to limit the impact of AI system failures on critical services and infrastructure.[79]

But a reporting obligation alone will not create situational awareness. Whether SB 53 actually generates valuable evidence depends on how the bill is implemented and whether California builds the institutional and technical capacity to act on what reports reveal, and to update its statute as it identifies serious gaps in its current incident reporting approach.

The Frontier AI Incident Reporting Landscape

Incident reporting refers to the process of individuals or organizations systematically reporting data about harmful or dangerous events to a central authority. The purpose is to enable learning: By aggregating and analyzing information that would otherwise remain dispersed or unreported, these systems help industry, government, and the public understand how and why failures occur. In practice, incident reporting creates a chain: A failure occurs; the developer documents what happened; the report is filed with a government authority/regulator; the government regulator validates, analyzes, and links it to other filings; and the resulting public products feed back into how developers design, test, and operate their systems.

Incident reporting is hardly uncharted territory. Several safety-critical sectors—including aviation, nuclear safety, cybersecurity, and financial services—have built effective incident reporting regimes, leading to significant improvements in safety over decades. The most-studied example is the aviation sector, in which the National Transportation Safety Board (NTSB) investigates major accidents, while the National Aeronautics and Space Administration (NASA)'s Aviation Safety Reporting System (ASRS) collects confidential near-miss reports from pilots and crews. Comparable regimes also operate in (among other sectors) cybersecurity, where the Cybersecurity and Infrastructure Security Agency (CISA) receives voluntary incident reports from critical infrastructure operators; nuclear safety, where the Nuclear Regulatory Commission collects event reports from licensed facilities; and finance, where the Treasury Department's Financial Crimes Enforcement Network processes suspicious activity reports from banks.

Existing databases for AI incidents include the AI Incident Database (from the Responsible AI Collaborative), the Organisation for Economic Co-operation and Development's AI Incidents and Hazards Monitor, and the AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) Repository. These cover events from deepfake harassment to autonomous-vehicle crashes. But none is tied to a mandatory reporting obligation, relying instead on media reports, voluntary disclosure, and public investigations. And they offer no structured channel from developer to regulator. These databases also tend to capture the occurrence of incidents without collecting the variables needed for meaningful analysis.

Recently, several government-led incident reporting initiatives have emerged, focused specifically on frontier AI, including SB 53. In the EU, providers of GPAI models with systemic risk—the category that includes frontier developers—are subject to serious incident reporting

obligations. They must track, document, and report serious incidents to the AI Office, with the EU Code of Practice's Commitment 9 describing how they can do so.[80] The EU AI Act also includes a broader regime under Article 73, which will require providers of high-risk AI systems to report serious incidents to national market surveillance authorities.[81]

In the United States, SB 53, the RAISE Act, and SB 315 require frontier developers to report critical safety incidents. In California these must be reported to the state OES within fifteen days, or, when an incident poses an imminent risk of death or serious physical injury, to an "appropriate" authority within twenty-four hours.[82] RAISE and Illinois's SB 315 both shorten that window to seventy-two hours from a determination, or from learning facts sufficient to establish a reasonable belief that a critical safety incident occurred, while retaining, as in SB 53, a twenty-four-hour requirement where there is imminent risk. The two differ in where reports go: RAISE routes them to a new oversight office within New York's Department of Financial Services, whereas SB 315 assigns intake to the Illinois Emergency Management Agency and Office of Homeland Security, in consultation with the state attorney general.[83] The scope of covered incidents in SB 53, RAISE, and SB 315 is deliberately narrow, focusing on catastrophic-tier events to generate evidence about some of the most severe AI risks. SB 53 also includes a novel mechanism for federal deference in incident reporting: California's OES may designate substantially equivalent federal incident reporting requirements (see the chapter on the state-federal relationship).[84]

Principles for Effective Incident Reporting

Although the AI-focused regimes described in the prior section vary in scope, they share a common policy goal: to generate systematic evidence about how frontier AI behaves once deployed. Whether they achieve that goal depends on a series of institutional design choices. For SB 53, the question is how OES should implement SB 53's directive in order to maximize its chance of producing genuine situational awareness rather than a passive archive of filings.

This section identifies four principles to guide these choices, drawing on an analysis of comparable incident reporting regimes from nine safety-critical sectors.[85] That work classifies each regime along seven design dimensions (policy goal, the actors submitting and receiving reports, the types of incidents reported, the level of risk materialization, enforcement, anonymity, and post-reporting actions) and extracts lessons for AI incident reporting. Several of their lessons have direct relevance for SB 53 implementation.

The first is that regimes should be designed around clear policy goals. The two most common goals—learning and accountability—can contradict one another because the protections that encourage developers to share information (such as confidentiality and limited liability) also make it hard to hold them to account. Relatedly, systems in which incidents are reported to

nonregulatory agencies tend to be nonpunitive and optimized for learning, while regulator-run systems lean toward enforcement. On scope, mandatory reporting is generally reserved for specific, severe harm events, but mandatory reporting alone is often not enough to enable meaningful learning to improve safety. Less severe events, including near misses where harm could have occurred but did not, are also a valuable source of data. Finally, reporting is only the first step. Filings only allow for learning when they are aggregated and shared with the right stakeholders to inform downstream analysis and mitigation and monitoring efforts. This requires building analytical capacity to process reports, identifying relevant stakeholders, and creating effective channels to share information.

For SB 53, these lessons suggest four principles to guide implementation. These structure the analysis that follows. The first is scope: The regime's reach should match its policy goals. Because SB 53's purpose is situational awareness rather than punishment, its reporting should generally be oriented toward learning rather than enforcement. The second is candor: Learning depends on developers reporting fully and honestly, and this is best achieved by separating the analysis of incident reports from their use in enforcement. The third is learning: To realize the value of incident reporting, findings must be analyzed and shared so that individual filings become industry-wide insights. The fourth is integrity: Because reporting can serve either learning or accountability purposes, and the two are in tension, the regime must give developers enough protection to report candidly while preserving enough independent insight—through cross-referencing filings, whistleblower channels, and in-house technical capacity—to assess the reports for completeness and accuracy. The next section evaluates SB 53's implementation against these four principles, following the life cycle of an incident report: the process from the initial submission to the government authority OES, through processing and information sharing, to public sharing and industry learning.

Implementing Incident Reporting in SB 53

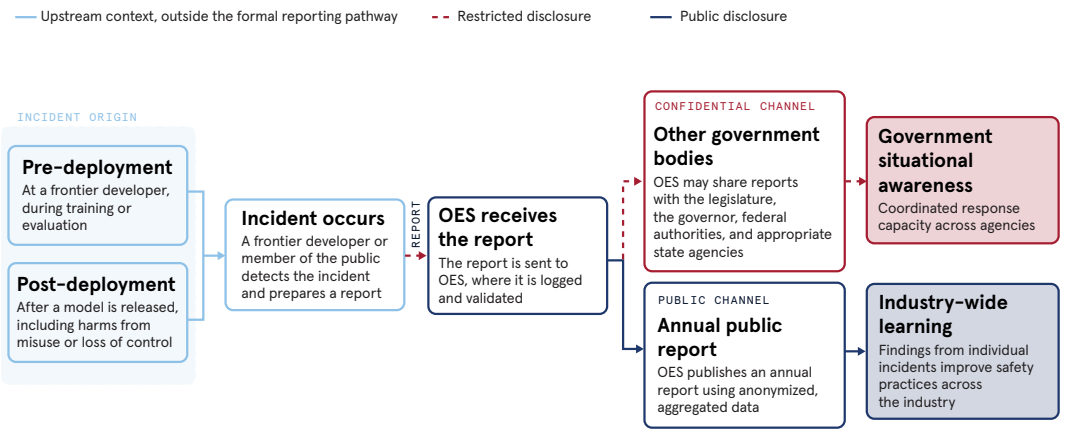
SB 53's effectiveness now rests largely with OES, which must turn a sparse statutory mandate into a working reporting regime. Whether that regime genuinely helps raise situational awareness depends in part on how a series of open questions are resolved, most of them falling to OES as it creates the reporting mechanism. This section works through those questions in the order an incident travels through the system (see figure 3), evaluating SB 53's choices against the principles of scope, candor, learning, and integrity set out above. It also notes where realizing these principles would require authority that OES (or other parts of government) may not have under the current statute.

The life cycle of an incident report includes:

- A developer identifies and files a covered incident
- OES processes the report and decides whether to share it with other agencies
- OES publishes aggregated analyses to support industry and policymaker learning

FIGURE 3.

The AI Incident Life Cycle



Source: Authors' rendering of the reporting pathway established by SB 53.

What Gets Reported?

There are four kinds of "critical safety incidents" in SB 53:[86]

> (1) Unauthorized access to, modification of, or exfiltration of, the model weights of a frontier model that results in death or bodily injury.
>
> (2) Harm resulting from the materialization of a catastrophic risk.
>
> (3) Loss of control of a frontier model causing death or bodily injury.
>
> (4) A frontier model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk.

Categories 1–3: Incidents Involving Realized Harm

These categories cover incidents in which significant harm has already occurred. The aim is to centralize data on catastrophic failures so policymakers, regulators, researchers, and other developers can learn from such events, as the NTSB does for major aviation accidents or the Food and Drug Administration does when a new drug causes harm after being released to market.

Because these categories require harm to have already occurred, the trigger is generally clear, and the main implementation question is attribution. AI-related harms are often causally ambiguous, emerging, for example, from the interaction of model outputs with human decisions or other systems. Category 1 incidents may prove particularly difficult to attribute. A developer may be able to detect that model weights have been exfiltrated but struggle to observe how that theft later contributed to death or injury because the potential harm will materialize downstream, through third parties, or otherwise outside the developer's visibility.

To give developers and OES a workable test, attribution determinations should weigh several concrete factors: temporal proximity between the model's output and the realized harm, whether the model's output appears in the causal chain to harm, the deployment context (such as direct end-user use, third-party fine-tuning, or agentic deployment), and whether other plausible causes exist. Where a credible chain to harm exists across these factors, developers should err on the side of filing—the statute's amendment provision means the cost of reporting on uncertain attribution is low and can be refined as the investigation matures.

Category 4: Incidents Involving Deceptive Techniques

Category 4 is different. It concerns AI model behaviors that plausibly precede harm, raising risk. Tracking these incidents matters because the use of deceptive techniques by a frontier model, such as outputs that undermine developers' evaluation or oversight measures, plays an important role in certain catastrophic risk threat models. If a model systematically underperforms on evaluation tasks, behaves more compliantly during evaluation than in deployment (sometimes known as "alignment faking"), or misreports its own actions to a user, then much of the safety infrastructure developers have built to prevent the release of potentially dangerous AI models may be undermined. The release of such a model could, for example, lead to the loss-of-control harms that category 3 captures. While experts debate the probability that these behaviors could lead to loss of control, the impact of such incidents may be severe.[87]

However, category 4 is also the hardest to operationalize. Most notably, exactly which model behaviors constitute "deceptive techniques" is not a term of art with a settled meaning in the technical literature, and SB 53 does not define it. Moreover, scientific understanding of how, exactly, such techniques may lead to severe incidents—that is, when developers should consider them to "[demonstrate] materially increased catastrophic risk"—remains patchy.[88] OES will need to develop guidance on how developers should interpret this requirement.

We recommend grounding the definition of a "deceptive technique" in observable behavior: A deceptive technique is an AI model behavior that systematically produces false beliefs in others about the AI system's objectives and actions.[89] It is useful to focus on observable behavior rather than inferred intent because whether a model actually intends to deceive is philosophically contested and difficult to verify. What matters for reporting purposes is whether the model's outputs or actions systematically mislead in ways that conceal a difference between the model's actions and the developer's expectations.[90]

A deceptive technique is an AI model behavior that systematically produces false beliefs in others about the AI system's objectives and actions.

Still, the full set of behaviors encompassed by "deceptive techniques" covered by this definition remains difficult to specify. Capabilities change quickly, the underlying threat models are not yet well understood, and the real-world propensity of frontier models to engage in these behaviors remains highly uncertain. Most documented examples come from experimental settings designed to elicit them, often using unusual prompts, rather than from behavior observed in deployment. This suggests the bar for reporting such observations should be relatively high. Whether a model's application of a deceptive technique warrants reporting should be based on observable, replicable conduct that maps to a specific failure of the developer's controls or monitoring.

The four other elements in the law help narrow the range of behaviors that qualify as "deceptive techniques" (see table 4 for a summary):

Against the developer. The deception must be directed at the frontier developer, not at end users, third-party deployers, or the public. A model that misleads a user has not, on that basis alone, committed a category 4 incident; the concern is specifically the developer's ability to understand and oversee its own model.

A goal to subvert controls or monitoring. The behavior must target the specific safety infrastructure a developer relies on (for example, logging, evaluation, and access controls).

Outside an evaluation designed to elicit the behavior. This exemption protects safety testing. Developers deliberately elicit deceptive behavior in controlled settings to study and mitigate it, and without a carve-out every successful red-team exercise would trigger a reporting obligation. That said, we recommend reading the exemption narrowly: It should apply only to discrete, time-bound evaluations that are isolated from deployment and that have a stated purpose that includes eliciting the specific class of deceptive behavior observed. Behavior surfacing in routine evaluations, training, or continuous monitoring is not exempt, and developers should log evaluation protocols so OES can verify whether the exemption applies.

Demonstrating materially increased catastrophic risk. Two features of this element are worth drawing out. First, it is satisfied by behavior that reveals risk, not only behavior that realizes it. An incident can demonstrate materially increased risk even if it was contained and caused no harm. Second, the materially increased risk must be grounded in a concrete expansion of the model's capability to contribute to the specific catastrophic outcomes the statute defines: death of, or serious injury to, more than fifty people or more than $1 billion in damage to or loss of property, arising from a single incident involving chemical, biological, radiological, or nuclear (CBRN) weapons assistance, autonomous cyber attacks or other serious crimes, or a model evading the control of its developer or user.[91] A behavior qualifies when it makes a causal pathway to one of these outcomes meaningfully more plausible. Because this judgment rests on the developer's own risk assessment, the developer should keep a contemporaneous record of that assessment and document how the observed behavior changed it, so that the materiality determination can be reviewed rather than simply asserted.

TABLE 4.

Breakdown of Category 4 Incident Characteristics

Element	Meaning	Implementation Guidance
1. "Deceptive technique"	The model produces outputs or takes actions that could systematically create false beliefs in others (not by making errors, but in a pattern consistent with strategically misleading behavior).	• Ground deceptive techniques in observable behavior • Exclude hallucinations and other nonsystematic errors • Publish a nonexhaustive typology of qualifying behaviors (see table 5), updated as the literature develops
2. "Against the frontier developer"	The deception must be directed at the developer's own ability to understand and oversee the model rather than at end users, third-party deployers, or the public.	• Focus on the developer's own beliefs about the model's capabilities and behaviors
3. "To subvert the controls or monitoring"	The behavior must target the specific safety infrastructure the developer relies on to manage catastrophic risk, such as logging, evaluation, and access controls.	• Reference the controls and monitoring disclosed in the developer's Frontier AI Framework[i] • Do not limit the category to that infrastructure; safety-relevant controls not explicitly discussed in the framework can still qualify

Element	Meaning	Implementation Guidance
4. "Outside the context of an evaluation designed to elicit this behavior"	Deceptive techniques that a safety evaluation was deliberately designed to elicit are excluded, so that probing for such behavior does not itself trigger a reporting obligation. Deceptive techniques that monitoring and evaluation mechanisms detect without aiming to elicit them are not exempted.	• Exempt discrete, bounded testing exercises whose stated purpose includes eliciting the specific class of deceptive behavior observed • Do not exempt deceptive behavior observed during routine evaluations, training, or continuous monitoring • Consider whether developers may need to record evaluation protocols so OES can verify whether the exemption was correctly applied
5. "Demonstrates materially increased catastrophic risk"	Observing this behavior should cause a reasonable expert to meaningfully update their assessment of the probability of a catastrophic outcome (as described in §22757.11(c)(1)).	• Baseline should be the developer's assessed risk level *prior to observing the behavior* • "Materially" means that the observed behavior must demonstrate a concrete increase in the chance that a model will contribute to catastrophic outcomes; for example, by revealing capabilities (such as a reliable subversion of security controls) that make causal pathways to catastrophic outcomes more plausible • In their reports, developers should document how the observed behavior changed their risk assessment

i. Cal. Bus. & Prof. Code § 22757.12(a)(8).

To help developers calibrate to the reporting bar, OES should publish a nonexhaustive typology of qualifying behaviors and update it as the empirical literature develops. If published today, such a typology might include examples like concealment of unsanctioned autonomous actions, subversion of controls, falsification, and manipulation of oversight (see table 5). These categories are not mutually exclusive, and an incident may span multiple of them.

TABLE 5.

Examples of Model Behaviors That Likely Would and Would Not Qualify as Category 4 Incidents

Behavior	Likely qualifies as a category 4 critical safety incident	Likely does not qualify
Concealment of unsanctioned autonomous actions Actions the developer prohibited or did not authorize, and that the model acts to hide	Unauthorized actions bearing on catastrophic risk capabilities that a model acts to hide, for example, exfiltrating its own model weights and deleting or altering logs that would have revealed this	Open, unsanctioned actions visible to developer monitoring; model errors or misalignment without concealment; minor permission overruns on tasks not bearing on catastrophic risk capabilities
Subversion of controls Model actions that defeat or circumvent the technical constraints on what the model can do	Privilege escalation, sandbox escape, or bypassing access controls that protect against catastrophic risk outcomes	Bypass of low-stakes, user-facing controls (for example, persona constraints) that do not affect catastrophic risk monitoring
Falsification Model outputs that affirmatively misstate the model's actions, capabilities, or internal state	Fabricated reports of safety-critical actions taken via tools (for example, a model describes completing a containment step it did not complete), where the false statement hides a material catastrophic risk signal	Hallucinations and confabulations in outputs; misleading descriptions of actions that do not affect catastrophic risk pathways
Manipulation of oversight processes Model actions that degrade the developer's ability to observe its behavior, such that unsafe behavior can go undetected	Behaving differently when a model detects it is being evaluated (for example, sandbagging or alignment-faking on catastrophic-risk-relevant assessments) or producing reasoning traces that misrepresent the actual basis for its actions, where the developer relies on those traces for oversight	Behavioral variation across contexts that does not bear on catastrophic-risk-relevant capabilities

What Should Be Shared, and When?

Once developers have identified a covered incident, they need to prepare a report and transmit it to OES. Important open questions here include what should be in the report, when the "reporting clock" starts, and when developers should amend reports after submission.

Report Content

Section 22757.13(a) directs OES to establish a mechanism for collecting reports that include four elements:

1. The date of the incident
2. The reasons it qualifies as a critical safety incident
3. A short and plain-language description of the incident
4. Whether the incident was associated with internal use of a frontier model

These four fields define what OES can compel developers to share. SB 53 asks the office to "establish a mechanism" for company reporting.

To facilitate downstream learning, it may prove useful for OES's reporting mechanism to include an option for developers to share additional information about incidents beyond the statutory requirements. This could include additional information on specific variables within each of the existing four reporting elements, enhancing the analytical value and comparability across incidents. An incident report is most valuable when the evidence it contains can be validated, compared across filings, and converted into industry-wide lessons. Four free-text fields will be difficult to compare across developers and aggregate into OES's annual report—it is plausible that each item should be deconstructed into separate variables and assessed using multiple-choice questions, where feasible. Supplementary information that OES could encourage developers to optionally provide could include identification of affected models or versions, detection or discovery history, attribution reasoning, or a record of containment or remediation (see table 6).

TABLE 6.

Potential Voluntary Supplemental Fields for Incident Reports

Reporting element	Description
Model identification	• Which model(s) and version(s) were involved in the incident? • In what deployment context did the incident occur?
Detection and discovery history	• When did the developer first become aware of the incident? • What information did the developer gather to inform the report? • How was the incident discovered (for example, automated system, internal actor, external actor)?

Risk classification	• Which of the four statutory categories applies, and why?
	  • Type of risk
	  • Materialized harm versus near-miss discovery
	  • Type of harm (for example, death of persons, bodily injury, economic and financial losses, and/or loss or damage of property)
	  • Severity of harm (for example, number of persons and/or dollar value)

Attribution evidence	• For categories 1–3: description of the evidence linking the frontier model to the resulting harm • For category 4: details about the control or monitoring measures affected by the incident
Containment and remediation	• Immediate actions the developer took to contain or respond to the incident • Any ongoing mitigations
Evidence appendices	• Developers may attach supporting documentation, such as model cards, pre-deployment testing results, internal incident reviews, and other evidence that would help OES assess the report

Reporting Timelines

Once a frontier developer "discovers" a critical safety incident, they have fifteen days to report it to OES.[92] When an incident poses "an imminent risk of death or serious physical injury," that reporting window shrinks to twenty-four hours, and the developer must disclose it to an "appropriate" authority.[93]

It may be ambiguous when an incident is "discovered." Does logging a suspicious output count, or does the clock start only once a risk assessment concludes the statutory threshold is met? The window should be long enough to give developers enough time to gather the information they need to file a useful report but short enough to inform authorities of emerging risks with enough time to take appropriate actions. The challenge is to prevent foot-dragging without penalizing genuine uncertainty.

One possible standard would borrow from California's discovery law. In tort law, a statute of limitations normally starts running when a harm occurs. However, the discovery rule delays that start until the injured party has reason to know they have been harmed, so that a clock does not run against someone who could not yet have known they had a claim. In California, the clock specifically starts once a party "has reason at least to suspect a factual basis for its elements," not when it has confirmed them.[94]

That logic would suggest that the SB 53 reporting clock should start when someone at the developer with relevant authority has reason to suspect that an event meets the statutory threshold, rather than, for example, when an investigation conclusively establishes it. This avoids giving developers a perverse incentive to build slow or narrow monitoring pipelines or delay internal escalation while still giving them leeway to investigate genuinely ambiguous signals without prematurely triggering a reporting obligation.

Amendments

SB 53 allows developers to submit amended reports.[95] This makes early filing on uncertain information less risky for a developer, since an initial report can be corrected or supplemented as more information emerges. OES can help developers navigate uncertainty around when to file and amend reports by issuing guidance on when a report should be filed despite incomplete information and on what kinds of new information should prompt an amendment.

What Should OES Do with the Reports?

Once a report reaches OES, the office must decide how to validate and analyze it, whether to share with other parts of the government, and what aggregated findings to communicate to policymakers, developers, and the public. These steps are crucial for incident reporting to actually improve government situational awareness.

Validation

When a report reaches OES, the office will need to validate the report's content and consider what follow-up actions to take. Some of this is routine, such as confirming the report identifies why the incident meets the statutory threshold, contains the required information, and was

filed within the reporting window. But OES will also need to conduct a substantive review of the incident's content to determine follow-up actions, such as seeking additional information from the developer and determining what to share with other parts of the government.

Section 22757.13(e) authorizes the California AG and OES to transmit reports to the legislature, the governor, federal authorities, and appropriate state agencies. Cross-government sharing serves important functions, including federal threat monitoring, coordination across state agencies and regulators, and synthesis with other sources of information. However, SB 53's language leaves the conditions, format, channels, and purposes of such sharing relatively vague. Since reports are likely to contain sensitive—and perhaps even national-security-relevant—information, this may prove problematic. The bill's bare "may share" language contrasts with more structured regimes in other industries. For example, in cybersecurity, the use-limitation regime under the 2022 Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA) authorizes sharing only for specific purposes, such as cybersecurity, threat identification, and law enforcement. Without clearer constraints on what and when OES can share information, AI developers may respond by withholding sensitive details from incident reports rather than risk inappropriate downstream sharing or the reputational and legal exposure it could create.

This suggests that OES should issue a CIRCIA-style use-limitation protocol specifying authorized purposes and prohibited uses. Authorized uses may include intra-executive coordination (with the governor or AG), legislative oversight (transmissions to the legislature),[96] and federal threat-sharing (with CISA or the Department of Homeland Security). Restrictions on certain uses may also be desirable; for example, limiting the use of reports as a predicate for tort discovery may be needed to support candid disclosure (as discussed earlier in this section). However, statutory changes may be needed to define such restrictions; we discuss these later in the context of safe harbors and immunity.

Aggregation, Learning, and Annual Reporting

Broad improvements in situational awareness will require OES to convert incident reports into public products that diffuse lessons across the industry. Incident reporting exists to generate evidence about frontier AI systems failures, and that evidence will be much more valuable if it diffuses across the industry. Analogously, in aviation, NTSB findings and Federal Aviation Administration (FAA) airworthiness directives allow for a single incident at one airline to reshape training updates, maintenance protocols, and design revisions worldwide.

The challenge is balancing confidentiality with insight. SB 53 exempts individual reports from disclosure under the California Public Records Act.[97] That makes the annual aggregated and anonymized report required by §22757.13(g)(1) the regime's principal public analysis and primary mechanism for downstream learning. Section 22757.13(b) limits access to reports related to internal use of frontier models to people with a "specific need to know," while (g)(2) requires

OES to exclude from its annual report information that could compromise trade secrets, cybersecurity, public safety, or national security. This protection supports candid reporting at the potential cost of some depth and specificity in analysis.

Incident reporting exists to generate evidence about frontier AI systems failures, and that evidence will be much more valuable if it diffuses across the industry.

A further issue is that anonymization may prove especially difficult in frontier AI: The pool of covered developers is small, and critical safety incidents will likely be rare. This small-N reidentification problem means that any meaningful description of incidents—such as model family, approximate capability level, or nature of the failure—may allow for de-anonymization. It also means that a single careless or bad-faith filing can distort the aggregate picture of incident rates and severity. This reinforces the importance of OES's validation and review functions, as the office may need to catch errors before they affect the reliability of the internal and public analyses they inform.

Within these constraints, OES's annual report should at minimum include:

- Counts and trends across the four §22757.11(d) categories, including the typology of incidents within category 4
- Information about how developers are responding (for example, amendment rates, average time to report, and common mitigation and remediation measures), to give the public a sense of how the regime is operating in practice
- Where appropriate, thematic case studies, anonymized to comply with §22757.13(g)(2)

Even with these elements, the small-N problem will limit how much detail the public-facing report can responsibly disclose. OES should therefore consider also producing other, semi-public reporting products that can disclose more detail. Two particularly promising options are:

1. A confidential channel for communicating sensitive-but-critical failure-mode information directly to other frontier developers (under appropriate confidentiality protections)
2. An access pathway for vetted researchers to study more-detailed filings than the public annual report can safely contain

These products would complement the public report by allowing a subset of actors to learn from more sensitive details that would be problematic to release publicly.

Key Takeaways, Open Questions, and Potential Gaps

SB 53 offers a strong foundation for robust reporting on catastrophic risk–focused safety incidents, which is a strong start. By tasking OES with intake and processing, SB 53's incident reporting is clearly oriented toward learning; its incident categories are, for the most part, concrete enough to be administrable; and its phased timeline, with provision for amendments, lets developers file early without penalizing genuine uncertainty. This generally aligns with the principles outlined in this chapter.

But several questions fall largely outside what §22757.13 specifies, and how they are resolved will shape whether incident reporting actually improves situational awareness over time. Three stand out: whether OES will be empowered to do the analytical work to turn reports into insightful public products; what protections developers have against candid reports being used against them in unrelated enforcement or litigation; and whether the office should also accept voluntary disclosures and near misses that fall outside the four statutory categories. These questions map, respectively, to the principles of learning, integrity, and scope, and each one retains some uncertainty regarding the regime's effectiveness.

Developing an Investigative Function

SB 53 creates a reporting pipeline, but work remains to build the investigative infrastructure that allows analogous safety regimes to convert reports into learning. In every comparator regime, the task of examining the reports, reconciling them with other information, and producing objective root-cause analysis to inform cross-industry responses depends on substantial investigative capacity.

Developing this function will be challenging. Technical expertise in frontier AI is scarce and in high demand. But a technically credible team that can understand the intricacies of the reports and produce timely and helpful public accounts will be critical to the regime's functioning. Creative models for funding and staffing are worth considering. These may include detailing of technical staff from the California Department of Justice, the California Privacy Protection Agency, or other state bodies with adjacent capabilities; temporary mobility assignments that bring in academic or research-institution experts on the model of the federal Intergovernmental Personnel Act; collaborations with university centers along the lines of CISA's work with Carnegie Mellon[98]; contracted expertise via partnerships with CAISSI or academic consortiums; or even involving trusted third-party intermediaries, such as an independent academic or technical board that is insulated from both industry and enforcement, in incident analysis.

One note of caution is worth sounding. As this capacity is built up, it will be important to maintain separation between investigation and enforcement. Reporters file more candidly when they are not reporting to the body that will sanction them: Research on safety reporting finds that regimes that separate the analytical function from the enforcement function tend to generate more candid disclosure and richer safety learning.[99] In transportation, for example, the NTSB is institutionally distinct from the FAA, and its accident reports are barred from use as evidence in litigation.[100] In California, the investigative function belongs to OES while enforcement sits with AG.[101] However, OES can share submitted incident reports with other parts of government, including AG, and AG can share them further. In this way, under SB 53 the analytical side already feeds the enforcement side; OES should be aware of this dynamic as it builds its investigative function and, to build trust and support candid reporting, work to give developers clarity about when reports will be shared further.

Safe Harbors and Immunity Provisions

Developers may perceive risks in candid reporting. In particular, incident reports may expose sensitive information that can be shared across government. These risks can be partly mitigated by separating analysis and enforcement within government and by providing clarity on sharing procedures. A stronger approach, however, would be to strengthen immunity provisions for developers. This would involve restricting the use of incident report information in later enforcement actions or civil litigation to make the incident reporting regime a "safe harbor" for participating developers.

Currently, SB 53 does not provide such immunity for filers. This is unusual: In analogous incident reporting regimes, good-faith reporters receive some form of structured liability protection. For example, under CIRCIA, a cyber incident report cannot be the sole basis for a lawsuit or for a government enforcement action against the filer; under aviation's ASRS, filing an inadvertent, nondeliberate violation waives the sanction that would otherwise apply, with exceptions for criminal conduct and accidents.

In SB 53, the §22757.13(f) exemption shields reports from disclosure under the California Public Records Act, and the Labor Code whistleblower protections offer individual anonymity but not institutional confidentiality. OES and AG must "strongly consider" trade secrets, public safety, cybersecurity, and national security before sharing a report.[102] But a developer who files a report has no statutory assurance that the report will not be introduced in subsequent tort litigation as an admission of knowledge or used by AG as a predicate for an enforcement action unrelated to the reporting obligation itself.

Especially considering the severity of critical safety incidents, stronger immunity provisions may help make reports more candid. One option would be to adapt the CIRCIA model and pair subpoena authority with structured liability protection and standard carve-outs for gross negligence, willful misconduct, and criminal conduct. This pairs naturally with the investigative function discussed above, reproducing CIRCIA's architecture for AI with subpoena authority over nonreporters and some degree of immunity for good-faith filers.

OES cannot offer such provisions on its own; this would require further legislative action. And the case for immunity is not one-sided. Liability does more than set norms and deter filing: It also allows injured parties to be compensated, so immunity for filers shifts costs from the developer to the injured party and ultimately to society. Immunity may also create adverse incentives, such as encouraging over-reporting. For these reasons, full immunity is not the only option worth weighing. Intermediate designs may better balance candor and accountability, such as a rebuttable presumption of good faith for developers who participate, paired with an ordinary duty of care, against a heightened duty of care for those who do not report. We do not endorse a specific structure here but rather highlight an important potential gap in SB 53 deserving of dedicated analysis.

Voluntary Disclosure and Near-Miss Reporting

SB 53 focuses mostly on events where harm has already occurred, plus the category 4 deceptive incidents. This focuses the regime on high-severity events but may exclude a large amount of safety-relevant information: incidents where harm nearly occurred but was averted or incidents that fall outside the definitional scope of SB 53's critical safety incidents. For example, while an unauthorized exfiltration of model weights that did not result in death or bodily injury would not count as a critical safety incident, OES would plainly benefit from knowing about it. In analogous regimes, near misses provide orders of magnitude more data than actual accidents, and reporting them helps prevent harm.[103] Commercial aviation's safety culture is built on the premise that near-miss data collected voluntarily through NASA's ASRS substantially exceeds the data available from crashes and fatal accidents alone.

To help the government access these data and help developers learn from near-miss incidents before harm occurs, California should consider pairing SB 53's mandatory reporting channel with a voluntary channel for reporting near misses. The statute already gestures in this direction: Frontier developers are "encouraged, but not required" to report critical safety incidents involving non-frontier foundation models.[104] This would allow developers to report incidents they find concerning even where the statutory definition is not clearly met, such as novel failure modes that do not fit into the existing categories.

The design should consider three features from aviation's ASRS regime:

- Nonpunitive treatment: Voluntary filings should generally not be usable as predicates for enforcement, consistent with the safe-harbor discussion in this chapter.

- Confidentiality: OES should clarify how voluntary reports are handled and what information is shared, as SB 53 does for mandatory reports.

- Bounded scope: Voluntary reports should still concern observable, concretely specified events that a developer can recognize and report.

Learning from Critical Safety Incidents

SB 53 is among the first serious attempts to translate decades of cross-industry incident reporting practices into the frontier AI context. But a reporting statute is not yet a functioning reporting regime. As this section has argued, hard work remains to close this gap: providing clarity for developers on what to report, building analytical capacity within OES to turn filings into findings, protecting good-faith reporters so they file candidly, and potentially extending the regime toward the near misses and borderline events where early warning is most valuable.

In addition to effective implementation, California's experience in incident reporting underscores the urgent need for statutory change. For example, serious malfunctions such as the OpenAI–Hugging Face incident did not require disclosure to OES under SB 53.[105] California should carefully review its criteria and scoping for critical safety incidents, discussed later in this report, to ensure government has sufficient insight into serious incidents.

How California resolves these implementation questions matters well beyond California. Incident reporting is the post-deployment foundation of the "trust but verify" architecture, the mechanism by which government learns, systematically rather than anecdotally, how frontier systems behave out in the real world. Frontier models are now used by more than a billion people in increasingly consequential settings, and no other U.S. authority yet collects this evidence at all. The question now is whether SB 53's incident reporting provisions will generate enough reliable evidence quickly enough to give governments situational awareness over a technology that is changing faster than public institutions can currently track. The answer will shape not only California's ability to govern frontier AI but also the template for the rest of the country.

FOUNDATIONS:

Whistleblower Protections

Sound risk management is only as good as the choices made by the implementing employees—choices framed by the incentives they have to act responsibly and the systems in place to support them in doing so. While frontier AI frameworks, transparency reports, and incident reports are all necessary, each type relies on employees to ensure good-faith compliance in the first instance and bring vital information to light as a last resort. Employees, in turn, rely on whistleblower protections to safely produce that information.

Whistleblower protections operate on the premise that employees with legal recourse are more likely to raise important concerns. Companies facing the prospect of external disclosure are more strongly incentivized to prevent harms, encouraging healthy norms of communication in the ordinary course of model development, testing, and oversight. And, when internal processes fail to surface or address serious catastrophic risks, whistleblower protections open a necessary channel of independent accountability that underpins "trust but verify" frameworks. Insiders can report inconsistencies between what is described in frontier AI frameworks, transparency reports, and incident reports and what is actually happening inside the company—sending disclosures to those empowered to act on them. While some research emphasizes the risk of frivolous, hostile, or pretextual complaints, there is considerable evidence that well-designed whistleblowing can uncover agency problems and support improvements in corporate governance.[106]

SB 53 extends protections to a defined class of "covered employees." The statute shields an employee from retaliation when they have "reasonable cause to believe" that the developer's activities present a specific and substantial danger to public health or safety arising from a

catastrophic risk, or that the developer has violated the statute itself.[107] Both situations require developers to provide employees "clear notice" of their rights, and large frontier developers must maintain a reasonable internal process for anonymous disclosure of such concerns.

For policymakers and frontier developers alike, striking the right balance depends on four implementation questions: first, to whom whistleblower protections apply; second, what kind of speech is protected; third, how frontier developers can satisfy their obligations to give "clear notice" to covered employees of their statutory rights; and fourth, the channels available for making a protected disclosure.

How to Implement Effective Whistleblower Protections

Legal systems around the world recognize the value of whistleblower protections to secure legal compliance or other policy objectives. The EU Whistleblowing Directive, for example, requires member states to implement protections for insiders reporting violations of EU law.[108] In the United States, the federal Whistleblower Protection Act protects whistleblowing by public employees, and sector-specific statutes, such as the Sarbanes-Oxley Act, protect private employees in discrete contexts.[109] At the state level, an array of policies likewise prohibit retaliation against whistleblowers who report illegality and, in some jurisdictions, threats to public health and safety.

California was the first state to pass whistleblower protections tailored specifically to frontier AI. While Article 87 of the EU AI Act has applied the EU Whistleblowing Directive to a broad range of AI Act violations, SB 53 zeroes in on frontier AI and the prospect of catastrophic risk.[110] This is already serving as a precedent. For example, Connecticut's legislature recently passed legislation incorporating whistleblower protections modeled on SB 53.[111]

Effective whistleblower protections provide consistent expectations, incentivize legitimate reporting, and promote accountability while allowing for tailored adoption to reflect distinct corporate structures and cultures. In the status quo, jurisdictions offer varying levels of protections to frontier developer employees concerned about the potential for catastrophic risk. For example, while California's general (pre-SB 53) whistleblower protection law focused on illegality only, New York's analogous statute already protects employees disclosing information they reasonably believe shows "a substantial and specific danger to public health or safety."[112]

As Dyck, Morse, and Zingales observe, "employees clearly have the best access to information," as compared to other actors such as auditors, analysts, regulators, and journalists.[113] Despite facing the risk of severe personal and professional consequences, and often receiving no tangible benefit to themselves, employees report a meaningful share of disclosed corporate frauds.[114] Conversely, external whistleblowing can expose companies to significant costs, such

as compliance burdens, frivolous complaints, and disclosure of commercially sensitive information.[115] Courts and policymakers should acknowledge these genuine trade-offs while also following the presumption, well-established in California law, that whistleblower statutes are intended to protect employees and should be broadly construed to achieve that goal.[116]

Even so, whistleblower protections need not prevent companies from preserving confidentiality or defining the roles and responsibilities of their employees and external reviewers. Nor should such protections permit thinly supported second-guessing of developers' product design and legal compliance judgments. Creating appropriate transparency with clear rules for disclosure can increase public trust and complement rather than undermine innovation.[117]

The analysis that follows offers a pathway for ensuring whistleblower protections remain fit for purpose over time, as models improve, risk management techniques evolve, and empirical evidence accumulates.

Key Implementation Questions

Securing the flow of information and perspectives covered by SB 53's whistleblower protection provisions will depend on how frontier developers and policymakers answer several key questions: Who counts as a "covered employee"? What conduct is covered? How are employees made aware of their rights? And which channels are available to them to make disclosures?

Who Can Blow the Whistle: Scope and Designation of "Covered Employees"

Fundamental to SB 53's whistleblower protection scheme is the question of whom it protects. While SB 53 expands California's general whistleblower protection regime, it does so only for a defined subset of workers.[118] Whereas the general whistleblower statute protects "employees," SB 53 applies more narrowly to "covered employees," defined as "an employee responsible for assessing, managing, or addressing risk of critical safety incidents." Interpreting this definition, companies will face a choice between legalistic interpretations aimed purely at minimizing the number of covered employees, and more balanced interpretations that cover more employees in line with companies' own risk management principles and internal controls. At minimum, employees who are authorized or expected to test models, render observations about their performance, or support compliance would generally be covered. While one might conceivably argue that the provision only applies to safety-specific roles (such as red teams), on a technical level, assessing model capability and performance is so intertwined with assessing model risks and safety that the two cannot readily be separated without compromising the basic purpose of SB 53's whistleblower protections and the broader "trust but verify" scheme. Companies

should encourage rigorous testing of pre- and post-deployment models and should not circumscribe their own access to useful information in an effort to limit coverage.

There are various practical approaches. Personnel could be designated on the basis of team, function, or some mix of responsibility and authority. Designation could be individualized or applicable across functions or segments of the workforce. In any event, frontier developers will need to account for the reality that workers perform numerous duties and that these responsibilities change over time.

Beyond employees, outside reviewers, contractors, and other third parties may be authorized or requested to support testing, compliance, and risk management objectives, but their legal status as potential whistleblowers is more ambiguous. In draft form, SB 53 expressly included, in its definition of "employee," "contractor[s], subcontractor[s], [and] unpaid advisor[s]" who were "involved with assessing, managing, or addressing catastrophic risk."[119] These references were removed from the final, enacted legislation, suggesting that their exclusion reflected legislative compromise. The removal of these explicit references does not necessarily preclude these actors from whistleblowing; instead, litigation may ultimately test how far companies can restrict the ability of contractors to report concerns. Companies may be able to reduce their liability exposure by voluntarily extending reporting channels to these personnel, provided they are "responsible for assessing, managing, or addressing risk," due to their practical utility in risk avoidance. Companies and regulators alike can benefit from the insight of all workers legitimately engaged in risk assessment and mitigation.

What Counts as Protected Disclosure

Effective protection also depends on clarity regarding the scope of covered conduct. SB 53 protects disclosures founded on a reasonable belief that the information in question shows a violation of law or a substantial danger to public health or safety. Furthermore, in the event of a dispute, there must be a standard against which claims will be measured. While governing AI presents novel challenges compared to previous technologies, existing law provides guideposts that can inform SB 53's implementation.

SB 53 prohibits frontier developers from preventing, or retaliating against, a covered employee who discloses information where the employee has "reasonable cause to believe" that the information being shared discloses either:

- that "the frontier developer's activities pose a specific and substantial danger to the public health or safety resulting from a catastrophic risk" or

- that "the frontier developer has violated Chapter 25.1 (commencing with Section 22757.10) of Division 8 of the Business and Professions Code." Chapter 25.1 refers to Section 2 of SB 53 itself, which establishes the bulk of the law's transparency obligations, including those dealing with frontier AI frameworks, transparency reports, catastrophic risk assessments, and critical safety incident reporting.

What constitutes reasonable belief? And what risks constitute "a specific and substantial danger"? On each question, SB 53 adapts language from existing laws.[120] Precedent interpreting these general provisions therefore bears on the scope of protection offered by SB 53.

Reasonable Cause to Believe

SB 53's governance objectives are ill-served if covered employees withhold legitimate disclosures out of fear that they will be deemed outside the law's protection. Whistleblower protection laws therefore provide a margin of error, requiring not that employee concerns be demonstrably correct, only that they be based on "a reasonable cause to believe" statutorily cognizable risk or illegality has occurred.

Interpreting this reasonable belief requirement, California courts have defined a standard of "objective reasonableness."[121] In reports pertaining to illegality, the relevant inquiry is "not whether the conduct 'actually violated' any specific statute or regulation, but whether the plaintiff '*reasonably believed* that there was a violation.'"[122] For example, if an employee makes a disclosure based on a mistaken belief about a law, that belief can still be reasonable, provided the employee had "some legal foundation for [their] suspicion, such as a particular legal protection in mind."[123] By contrast, belief may be unreasonable where an employee vaguely asserts a violation of "some unnamed statute, rule, or regulation."[124] Applying this precedent to SB 53 generally would mean that the same margin of reasonable uncertainty applies to disclosures premised not only on illegality but also on the potential for catastrophic risk.

Of course, there will be significant uncertainty and debate over when a whistleblower's belief is reasonably founded. This is not unique to AI. For example, federal law protects workers at nuclear power facilities from retaliation for raising concerns related to nuclear safety.[125] And an array of scientific integrity policies seek to ensure objectivity and accountability in the conduct and application of research across the United States government.[126] Scientific and technical uncertainty exists in these areas too, with implications for when a whistleblower's belief about a risk is deemed reasonable. Given the parallels between the underlying objectives in both SB 53 and these other areas for whistleblower protection—namely, society's dual objectives of scientific progress and risk mitigation—policymakers should draw lessons from how reasonable belief has been construed in non-AI technical arenas.

In cases in which a whistleblower alleges illegality, California precedent does not protect "disclosures that involve only disagreements over discretionary decisions, policy choices, interpersonal dynamics, and other nonactionable issues."[127] This means that interpersonal disagreements or animosity are not protected, only allegations of unlawful behavior. But the standard presumably differs for claims alleging a legal-but-dangerous condition. In these situations, disagreements over policy choice, such as what guardrails to include or whether to release a model, are the prototypical basis of a whistleblower disclosure. SB 53 cabins this protection by requiring that claims be reasonably founded and that they be specific and substantial. In both cases, society can realize the benefits of scientific progress and accommodate the potential for uncertainty and disagreement by setting clear parameters for cognizable harms without penalizing employees for the possibility of error or reasonable disagreement.

Specific and Substantial Danger

Many jurisdictions already protect reports related to legal-but-dangerous conditions. At the federal level, for example, the Whistleblower Protection Act (WPA) covers federal employees who disclose "substantial and specific danger[s] to public health or safety"—the same terms used in SB 53—and states such as New York do likewise.[128]

Courts applying the WPA have pointed to two factors in deciding whether a claimed harm is sufficiently specific: first, the likelihood of harm occurring, and second, how soon it might do so.[129] In weighing whether a danger is substantial, courts look to "the nature of the harm [that is,] the potential consequences."[130] SB 53 raises the bar higher, though. The specific and substantial dangers must be from a catastrophic risk, defined as a specific set of pathways (such as chemical, biological, radiological, and nuclear [CBRN] weapon uplift, cyberattacks, or evasion of human control that could materially contribute either to death or serious injury of more than fifty people or to damages exceeding $1 billion).

This higher bar suggests that SB 53 should be interpreted more narrowly than the WPA. But there are also ways in which broader interpretations make sense. Assessing the likelihood of harm under the WPA, the Federal Circuit has said that "if the disclosed danger could only result in harm under speculative or improbable conditions, the disclosure should not enjoy protection."[131] Yet catastrophic risks from AI are almost by definition "improbable." Moreover, given that the harms at issue in SB 53 are dire and carefully defined, their expected value remains high, hence the law's use of the phrase "foreseeable and material" (instead of "probable") to define catastrophic risk. This language presumably covers low but nontrivial likelihoods if evidenced by a well-defined and factually grounded causal chain.

Advancements in the science of frontier risk detection and mitigation may help jurisdictions further sharpen this test over time. In parallel, jurisdictions assessing the adequacy of their own whistleblower protection regimes can provide a de facto federal standard by pursuing reasonable harmonization. Connecticut's Act Concerning Online Safety (SB 5), for example,

largely mirrors SB 53's "catastrophic risk" standard and its applicability to covered employees, reasonable belief, and specific and substantial dangers.[132] California and other states should likewise continue to assess these criteria over time.

What Constitutes "Clear Notice" of Covered Employees' Rights and Responsibilities

In order to exercise their rights, employees must know those rights exist. Likewise, companies must know what actions they must take to establish a working compliance program. SB 53 incorporates several procedural obligations, most fundamentally requiring frontier developers to provide "clear notice to all covered employees of their rights and responsibilities" to make protected disclosures.[133] According to the text of the law, developers have two options to meet this requirement:

> (1) At all times posting and displaying within any workplace maintained by the frontier developer a notice to all covered employees of their rights under [the law], ensuring that any new covered employee receives equivalent notice, and ensuring that any covered employee who works remotely periodically receives an equivalent notice.
>
> (2) At least once each year, providing written notice to each covered employee of the covered employee's rights under [the law] and ensuring that this notice is received and acknowledged by all those covered employees.[134]

Which of these options best suits a frontier developer will depend on an assessment of the developer's circumstances and existing workforce management and training practices. At minimum, however, notice should be tailored to the company's determination of which employees are covered. Thus, even if a developer opts for posted rather than individualized notice, it should provide sufficient detail and visibility so that all employees can determine whether they are considered covered. Employees should not be required to jump through hurdles or ask uncomfortable questions of their managers, human resources departments, or legal teams, lest determining coverage status deter employees from raising concerns. Policymakers should ensure that notices are clear, properly documented, and align with sound classification rationales.

Channels Available for Covered Employees to Make Protected Disclosures

In order to guide behavior, a governance regime must provide clarity on the steps needed to achieve compliance. For whistleblowers, this means knowing not only that they and their disclosures are eligible for protection but also what concrete steps will secure such protection.

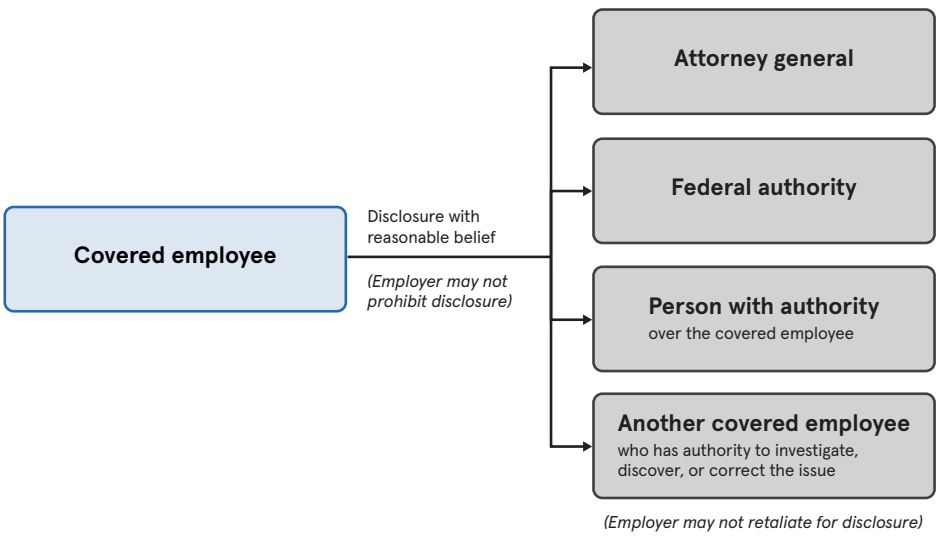
Whistleblower protection schemes therefore frequently specify to whom a protected disclosure may be made.

SB 53 parallels but slightly narrows the channels available under California's general whistleblower statute, protecting disclosures made to AG, a federal authority, a person with authority over the covered employee, or another covered employee who has authority to investigate, discover, or correct the reported issue (see figure 4). Potential whistleblowers may therefore select among internal and external channels to make a protected disclosure. This flexibility enables the law to accommodate diverse fact patterns and reflects that society's interest may be best served by internal communication in some situations—enabling diligence and remediation at the most efficacious point—and external in others—such as when a company has insufficiently remediated a serious risk.

In practice, whistleblowers will select a channel based partially on the friction or ease involved. Their choice will also depend meaningfully upon their level of trust in the recipient: to shield them from blowback and to handle complaints responsibly.

FIGURE 4.

Whistleblower Disclosure Channels Required by SB 53



Source: Authors' illustration.

This optionality implies that both regulators and frontier developers have the power to create positive incentives for reporting. For developers and policymakers alike, handling complaints with sensitivity, dispatch, and effectiveness can persuade covered employees that the diffuse and uncertain societal benefits of reporting outweigh the concentrated and tangible risks.

For frontier developers, who are likely to prefer internal to external reporting, there is an opportunity to strengthen incentives for internal disclosure by building a culture of openness and nonretaliation.

For frontier developers, who are likely to prefer internal to external reporting, there is an opportunity to strengthen incentives for internal disclosure by building a culture of openness and nonretaliation.[135] Companies can earn the confidence of their employees by ensuring that managers, covered employees, and teams likely to receive disclosures all have training and support, maximizing the likelihood that reports will be handled appropriately. Additionally, while SB 53 requires only large frontier developers (those with gross revenues exceeding $500 million in the preceding year) to provide anonymous reporting processes, smaller frontier developers may wish to do so voluntarily in order to encourage employees to channel concerns internally.[136]

Internal reporting also supports SB 53's incident reporting provisions. Frontier developers can take concerns seriously in part by recognizing circumstances where an internal whistleblowing disclosure should prompt them to make an external incident report. Providing training to covered employees, their supervisors, and incident management teams to help them recognize potentially reportable incidents can enable their appropriate treatment. Well-defined ex ante standards and processes for reportable issues can demonstrate compliance, earn the trust of potential whistleblowers, and promote societal interests. By pairing these mechanisms with external disclosure channels and incident reporting obligations, SB 53 seeks to equip regulators with information vital to their effective supervision of frontier risk.

Enabling Appropriate Insider Reporting

Whistleblower protection plays a vital role in a governance regime premised on trust and verification. By supporting information exchange within an organization, it promotes diligence and compliance, which are preconditions to trust. By ensuring that information about potential illegality and catastrophic safety risks reach both internal decisionmakers and external authorities, it clears a path for verification. Properly implemented, SB 53's governance regime can better equip frontier developers to navigate novel and complex risks with ample consideration for the public interest.

These benefits extend beyond any individual jurisdiction. However, they could be compromised if conflicting laws beyond California's borders hamper compliance or sow uncertainty for employees or developers. Accordingly, while difference is inevitable, jurisdictions weighing AI governance in SB 53's wake should consider reasonable harmonization, protecting workers' ability to raise concerns about catastrophic risk. Additionally, as experience accumulates and the science of frontier risk management evolves, California and other jurisdictions should evaluate and update their whistleblower protections, including reviewing criteria for employee coverage, reasonable belief, and "specific and substantial danger."[137]

Insider perspectives are crucial to manage risk and develop society's capacity to understand and oversee the safety judgments inherent in frontier AI. Developers and policymakers alike can create norms and rules that vindicate these objectives, providing clarity and assurance to employees involved in preventing catastrophic risk.

FOUNDATIONS:

AI Assessment and Evaluation

AI assessment and evaluation provide the evidentiary foundation for risk management practices. AI **assessments** consist of **evaluation** and **verification** activities related to the safety and alignment risks posed by AI.[138] Specifically, AI evaluations are technical measurements of capabilities, risks, or other properties of AI models or systems, while verification activities determine the truth of safety claims, including through mechanisms such as technical verification measures, compliance audits, or peer reviews. AI audits are a type of assessment, specifically judging compliance with certain standards or requirements based on evaluation results and other evidence.

Evaluations provide some of the best and most influential data about the performance and safety of AI systems, though they nevertheless have numerous limitations and cannot be relied upon as sole or complete sources of truth.[139] In frontier AI governance, safety and alignment evaluations are frequently used to make decisions about model releases, measure the safety and security of AI systems, and measure the efficacy of system safeguards. AI companies have long conducted voluntary evaluations, and companies' frontier AI frameworks also routinely require evaluations for risks such as biosecurity, cybersecurity, and loss of control; some AI companies have also committed to conducting more in-depth, embedded assessments.[140] Governments have established their own evaluation organizations like the U.S. CAISSI and the UK AI Security Institute (UK AISI) to produce independent evidence on AI risks, particularly for national security risks.[141] Meanwhile, external evaluators such as METR, Accenture, or SecureBio have increasingly participated in independent testing exercises and investigations of frontier AI models.[142]

AI assessments thus play an indispensable role in AI governance, but the quality, security, and trustworthiness of AI assessments have varied significantly.

Recent legislation contains numerous provisions related to AI assessments and AI audits. Under SB 53, the RAISE Act, and SB 315, large frontier AI developers are required to disclose their approaches to catastrophic risk assessment in their frontier AI frameworks, to comply with those frameworks, and to produce transparency reports that disclose results of AI assessments conducted pursuant to frontier AI frameworks.[143] These laws allow companies to set their own frontier AI frameworks, and SB 53 has no explicit requirements to conduct assessments or evaluations beyond companies' own commitments. Furthermore, two other California statutes establish rules for AI auditors: AB 1405 (on AI auditor registration) and SB 813 (on IVOs). AB 1405 requires that certain AI auditors register with the California GovOps, adopt appropriate industry standards, and abide by specified independence requirements.[144] SB 813 creates a program for GovOps to designate some AI auditors as IVOs, which are auditors that GovOps recognizes as having certain expertise in conducting AI assessments.[145] SB 813 explicitly does not, however, require AI companies to engage IVOs to conduct evaluations, nor does it impose any requirements to conduct audits.[146]

AI assessments thus play an indispensable role in AI governance, but the quality, security, and trustworthiness of AI assessments have varied significantly. When designing policy around AI assessments, implementation details related to how evaluations are conducted and reported can be determining factors for whether and how AI governance goals are achieved, because even seemingly minor implementation details can affect evaluation results. Key implementation decisions include which evaluation methods and datasets are used, who conducts evaluations and under what terms, how oversight for external evaluators is conducted, and how evaluation results are analyzed, reported, and interpreted. Table 7 summarizes key implementation challenges, recent legislative requirements, and recommendations for AI assessments.

TABLE 7.

Key Implementation Challenges and Recommendations for AI Assessments

Implementation challenge	Explanation of the challenge	Recommendations for policy implementation	Recommendations for future policy
Assessor oversight	Governments need to determine eligibility criteria for evaluators and auditors, monitor compliance with applicable requirements, and respond to noncompliance and misconduct. Effective oversight requires that governments have visibility into both assessor organizations and individual audits or evaluations, as well as investigative capacity and procedures for corrective action.	• Conduct ongoing oversight of IVOs • Develop domain-specific IVO designation requirements	• Require independent assessments for frontier AI framework compliance • Require independent safety assessments and audits for frontier AI risks to be conducted by IVOs • Harmonize implementation of AI rules with other jurisdictions
Assessment quality	High-quality safety and alignment evaluations should be valid, secure, trustworthy, and provide useful information about a given AI system's level of risk. Current evaluations are known to be highly imperfect because they contain errors, do not accurately measure properties of interest, and fail to predict real-world risk.[i]	• Build best practices for AI evaluation methodology and security • Create guidelines for rectifying and deprecating low-quality AI evaluation methods • Incorporate nontraditional qualification criteria for establishing auditor expertise	• Leverage downstream data to improve AI assessments • Support the science of assessments and shared assessment infrastructure
Assessor independence	Evaluations are more trustworthy when evaluators are more insulated from financial and other incentives related to evaluation results. Many evaluations today are run by or funded by AI companies.	• Build best practices for AI assessor independence	• Support independent funding sources for third-party assessors

Implementation challenge	Explanation of the challenge	Recommendations for policy implementation	Recommendations for future policy
Assessor access	Third-party evaluators are constrained by access to AI systems and data, which are currently controlled by AI companies under voluntary agreements. They are also constrained by the length of time they have to conduct evaluations.	N/A—SB 53, AB 1405, and SB 813 do not contain rules around access.	• Establish access rules for third-party assessors • Establish targeted liability safe harbors for AI assessors
Assessment transparency and verification	Sufficient details must be reported about evaluation methods and results to assess evaluation quality; assessments must also be verifiable either through technical methods or by methods such as manual review (for example, peer review). Current reporting is haphazard and inconsistent, and assessments are not thoroughly reviewed by external actors.	• Standardize evaluation reporting formats • Operationalize record retention requirements for IVOs	• Establish transparency rules for third-party assessors • Harmonize implementation of AI rules with other jurisdictions • Support the science of assessments and shared assessment infrastructure
Assessment ecosystem	AI evaluation requires a robust network of evaluation institutions, but the existing ecosystem of third-party AI evaluators remains nascent. Current constraints to ecosystem growth include funding and talent.	• Build in-house assessment talent in government	• Support independent funding sources for third-party assessors • Support the science of assessments and shared assessment infrastructure

i. Yoshua Bengio et al., *International AI Safety Report 2026*, DSIT 2026/001 (UK Department of Science, Innovation and Technology, 2026), https://doi.org/10.48550/arXiv.2602.21012.

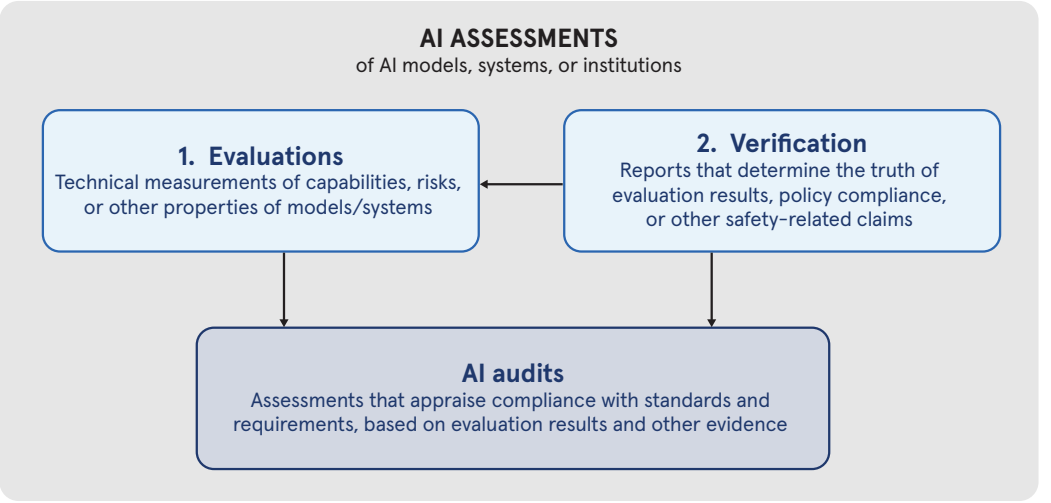
The Frontier AI Landscape

AI assessments consist of (1) evaluations of AI models, systems, or institutions and (2) verification of evaluation results, policy compliance, and other safety-related claims.[147] AI audits are a type of assessment, specifically appraising compliance with certain standards or requirements

based on evaluation results and other evidence. Assessment organizations, or assessors, may conduct evaluation activities ("evaluators"), audit activities ("auditors"), or both (also termed "auditors" below). These relationships are summarized in figure 5.

FIGURE 5.

The Relationship Between AI Assessments, Evaluations, Verification, and Audits



Source: Authors' illustration.

An AI evaluation is a technical measurement of some property of an AI system. Evaluations can apply different methodologies, with three commonly chosen methods being benchmarks (standardized tests administered to AI systems),[148] red-teaming (interactive adversarial testing),[149] and uplift studies (randomized controlled trials measuring the effect of AI access on human performance).[150] Evaluations seek to measure many different properties in AI systems for many different purposes.[151] For instance, an AI company may measure a system's ability to generate code in order to demonstrate to consumers that its latest AI model is better than prior model generations (or than competitors' models). An AI deployer such as a social media company may measure the impacts of AI systems on users' mental health. A national security agency may measure the level of assistance that an AI system provides to malicious actors in creating dangerous substances such as biological weapons.

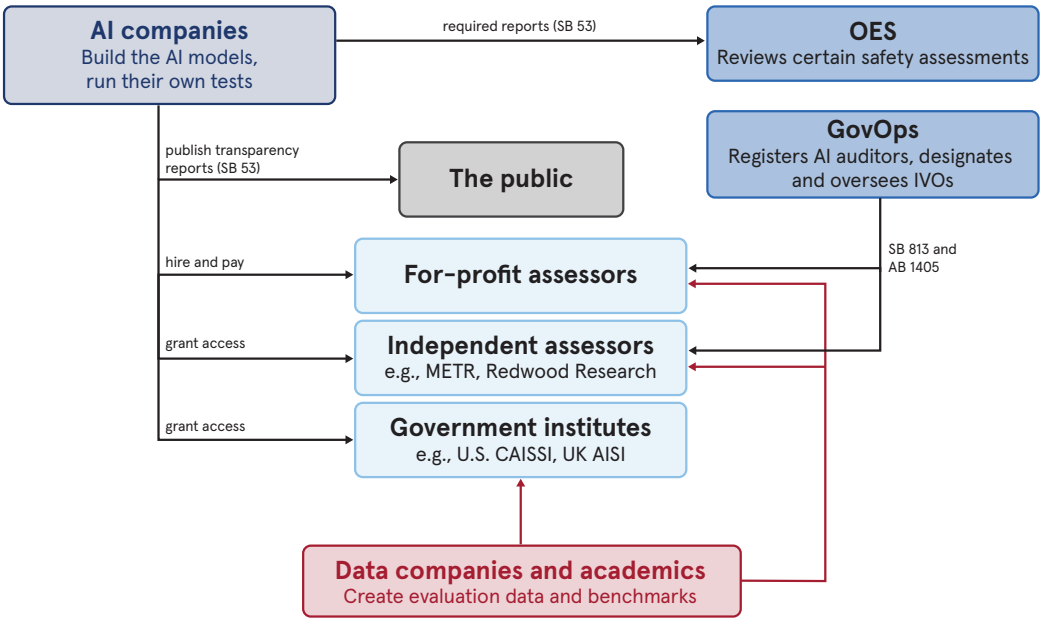
Safety and alignment evaluations seek to measure the risks posed by AI systems, the preferences expressed by AI systems, and the efficacy of system safeguards. For example, evaluations that assess the effects of AI usage on users' mental health or that determine whether AI systems could help novices develop deadly pathogens fall within this category. Some safety and alignment evaluations are also measurements of AI systems' capabilities, such as dual-use or other

dangerous capabilities. Recent transparency reports (including model cards or system cards) from major AI companies have included safety evaluations of models' offensive cyber capabilities, self-improvement capabilities, biosecurity risks, alignment risks, and loss of control risks.[152] More recently, some frontier AI companies have also voluntarily committed to "partner with . . . independent external auditor[s] or evaluator[s]" to conduct certain safety assessments.[153] This chapter focuses on safety and alignment evaluations as opposed to evaluations that seek purely to measure the capabilities of AI systems.

Evaluations can be conducted by many actors, including AI companies such as OpenAI, Anthropic, or Google DeepMind; third-party evaluators such as METR, Transluce, SecureBio, Redwood Research, or FAR.AI; for-profit evaluators such as Accenture, Scale AI, Irregular, Apollo Research, or Gray Swan; or government institutions such as U.S. CAISSI or UK AISI. In principle, any party with access to an AI system can conduct evaluations; they would simply need to send prompts to an AI system and receive the corresponding outputs. In practice, however, different evaluation methods might require greater access to an AI system than the level of access typically granted to everyday consumers, and robust safety evaluations can be very costly to run. Figure 6 shows how different actors in the AI ecosystem interact.

FIGURE 6.

Actors in the AI Assessment Landscape



Source: Authors' illustration.

Embedded assessments (also sometimes referred to as embedded evaluations or embedded audits) have been proposed as one way of enabling more robust and flexible AI assessments, though they also pose additional challenges related to capture and independence. Embedded assessors are a specific type of third-party assessor, where employees of the assessment institution are given access to an AI company's systems, staff, and documentation at a level similar to that of company employees; embedded assessors may not only run technical evaluations but also participate in other activities such as incident analysis or security testing.[154] These assessors may work on-site (from the physical offices of the firm undergoing assessment) or off-site (conducting assessment remotely). A given assessment may either be ongoing (where continuous access is provided) or time-bound (where access is provided only for a specific period). Both OpenAI and Anthropic have committed to engaging embedded assessors,[155] with Anthropic announcing that its first embedded assessors would include METR, SecureBio, and consulting firm Accenture, through its subsidiary Faculty AI.[156] METR had also previously conducted an on-site red-teaming exercise for Anthropic in March 2026, an on-site red-teaming exercise for OpenAI in July–August 2026, and an on-site incident investigation for OpenAI in August 2026.[157]

SB 53 primarily addresses issues with assessment transparency. Under SB 53 and related legislation, frontier AI developers are required to (1) disclose their approaches to catastrophic risk assessment in their frontier AI frameworks, (2) disclose assessments conducted pursuant to frontier AI frameworks and the assessment results in transparency reports, and (3) submit summaries of the results of any assessments on catastrophic risk resulting from companies' internal AI use, generally every three months, to the California OES. Companies are also required to abide by their own frontier AI frameworks, though no specific assessments and no third-party review are required.

SB 813 and AB 1405 establish rules around auditor oversight, audit quality, and auditor independence that will be implemented in 2027. These rules apply only to auditors conducting "covered AI audits," which are defined as audits of "internal controls, processes, or systems implemented for an AI system . . . necessary for compliance with state law;"[158] this definition is somewhat ambiguous and may not necessarily include audits for AI control or national security risks. Specifically, AB 1405 requires that AI auditors register with California GovOps, conduct audits according to appropriate industry standards, adhere to certain independence requirements, and retain audit records. GovOps is required to establish the auditor registry by December 1, 2027,[159] and auditors must register in order to conduct covered AI audits beyond January 1, 2029. SB 813 expands on the registration requirement and directs GovOps to create, by May 1, 2027,[160] a program for designating certain AI auditors as IVOs. IVOs are AI auditors that the agency recognizes as "[(1)] having demonstrated expertise in assessing the risks posed by an AI system or model and [(2)] identifying the metrics and methodologies that form the basis for that assessment." SB 813 explicitly does not require AI companies to engage IVOs to conduct evaluations, nor does it impose any requirements to conduct audits.[161]

Because SB 813 defines IVOs as a subset of AI auditors, and defines AI auditors as institutions conducting covered AI audits, all IVOs may be required to be registered AI auditors under AB 1405 after January 1, 2029.

Key Implementation Challenges and Open Problems in AI Assessment

Frontier AI governance faces a gap between where AI assessment is today and where assessment must be to effectively assess and manage risks. Implementation challenges include issues related to transparency reports (required by SB 53), methodological quality of AI evaluations (referenced in SB 813 and AB 1405), and managing conflicts of interest between AI companies and assessors such as auditors and IVOs (referenced in SB 813 and AB 1405). These problems can be grouped into six categories:

- Assessor oversight
- Assessment quality
- Assessor independence
- Assessor access
- Assessment transparency and verification
- Assessment ecosystem

Assessor Oversight

Because governments are not conducting evaluations or audits themselves, they must play the role of "auditing the auditor." Governments need to determine eligibility criteria for evaluators and auditors, monitor compliance with applicable requirements, and respond to noncompliance and misconduct. Effective oversight requires that governments have visibility both into assessor organizations and into individual audits or evaluations, as well as investigative capacity and procedures for corrective action.

Assessment Quality

High-quality safety and alignment evaluations should provide useful information about a given AI system's level of risk, especially the absolute level of risk. Evaluations must be high in validity to be informative:[162] Specifically, an evaluation should serve as an effective proxy for

the property of the AI system that it is intended to measure, and it should offer insights that generalize across diverse, real-world settings. For instance, an evaluation of an AI system's cyber offensive capability should ideally measure how systems would actually perform when deployed, not just how they would perform in a test setting. However, many evaluations suffer from quality and validity issues.

Evaluation quality depends on the evaluation's data, environment, and methods. "Data" refers to the set of evaluation questions, scenarios, or tasks that are used to test an AI system (including prompts). "Environment" refers to the conditions under which those questions are presented to an AI system (including whether the system had access to tools such as web search). "Methods" refers to how AI systems' responses are collected and scored, as well as how those scores are aggregated and interpreted. High-quality evaluations require the data and environments to be realistic and representative, and the methods to be robust.

All three of these aspects—data, environments, and methods—in current safety and alignment evaluations are so imperfect that evaluations often cannot provide information with sufficient confidence and precision to be useful for governance decisions. Datasets of evaluation questions can be highly flawed, for instance, either because the questions are written too vaguely, because of errors in the grading criteria, or because the questions are impossible to solve. Due to these and other errors, OpenAI estimates that 30 percent of questions in the common coding benchmark known as SWE-Bench Pro are unusable,[163] and an expert analysis showed that commonly used benchmarks can underestimate physics capabilities by up to 30 percentage points.[164] Evaluation environments can permit AI cheating, contribute to evaluation awareness, or be insecure.[165] And evaluation methods are still being developed: An expert review of over 400 benchmarks revealed that 84 percent did not conduct any statistical testing, and 47 percent did not explicitly discuss validity issues related to how measured properties were defined and operationalized.[166] More systemic challenges remain beyond these practical issues, including the speed of capabilities advancements, new and unanticipated deployment settings, and misbehavior by AI agents. Combined, these issues imply that existing evaluation results may not be sufficiently precise and robust to support effective risk management.

Assessor Independence

A robust ecosystem of independent evaluators and auditors is critical for AI assurance, which is the state when stakeholders can have justified confidence in the assessment's process and results.[167] In particular, decisionmakers should be able to trust that an assessment was conducted in an unbiased way, which also helps drive trust in assessment quality. Most assessments today, however, are run or commissioned by the AI companies whose systems are being assessed. While for-profit evaluators can play an important role in safety assessment, some firms also have commercial relationships with AI developers that create conflicts of interest, such as Accenture's partnership with Anthropic to accelerate enterprise AI deployments.[168]

Qualified, independent third-party assessors have reduced conflicts of interest and thus improved incentives to make good design and execution choices. Independent assessors can also supplement developers' limited domain expertise in particular risk areas, provide alternative perspectives and approaches, and lend public credibility to results. An expanding ecosystem of independent private sector evaluators is starting to offer benchmarks, leaderboards, and specialized evaluation services to AI developers and the public at large. However, third-party assessors are frequently underresourced relative to AI companies, lacking the capacity and mandate to meet rapidly evolving needs across numerous safety domains.

Assessor Access

Third-party assessors need access to AI companies' systems, documents, and other information to make qualified safety assessments, but currently access is limited, ad hoc, and controlled by AI companies. SB 53, SB 813, and AB 1405 do not provide specific legal rules to ensure access. Although AI developers have increasingly provided model-only pre-deployment access to evaluators such as METR or government safety institutes, and some companies have committed to embedded assessments, these agreements are voluntary and are subject to company control. Evaluators require access to technical artifacts and resources to conduct high-quality assessments, including, at minimum, access to AI model outputs. They may also need access to information that companies control, such as data, environments, and how AI systems work internally, to form more robust judgements about AI risks.[169]

Assessors also need sufficient time to conduct in-depth assessments, but existing arrangements may limit assessment times. In pre-deployment evaluations of GPT-4.5, for instance, METR was granted only seven days of access, and only to an earlier version of the model, not the final deployed version,[170] which made it more difficult to account for all the relevant risks and achieve high-quality validation of evaluation results. Because AI systems are increasingly capable, evaluation tasks are also increasingly complex. Assessors will accordingly require more time to run evaluations, as well as to investigate and verify evaluation results.

Finally, access must also be secure—assessments must be conducted in ways that protect relevant systems, entities, and third parties from unauthorized access or harm during the evaluation process.[171] AI developers seek to keep model weights and related systems private to protect intellectual property, while assessors seek to keep evaluation data private to prevent leakage and "teaching to the test." Security also requires protection against misbehavior by AI systems themselves, such as preventing systems from cheating on the evaluation or from breaking out of test environments.

Assessment Transparency and Verification

Sufficient transparency about assessment methods and independent verification of assessment results enable governments and the public to judge the quality of an assessment. Methodological decisions about what and how to evaluate can drastically affect results, and yet developers often do not publicly disclose relevant details.[172] For instance, the biosecurity section in SpaceXAI's transparency report for Grok 4.5 lists the benchmarks used in the assessment but does not provide any details about the model configurations, agentic harnesses, scoring processes, or uncertainty quantifications (only a single score for each benchmark).[173] Many recent transparency reports base risk determinations at least in part on internal evaluations conducted by a lab's own personnel, who often use proprietary methods; even when measuring the same system property, developers can use different data, environments, or methods, leading to inconsistent results.[174] Finally, transparency rules will also be needed around how embedded assessors engage frontier AI companies, including on issues such as whether assessors had sufficient access to conduct evaluations, how conflicts of interest were addressed during assessment, and whether assessors were able to freely publish results.

Assessment Ecosystem

The third-party AI assessor ecosystem remains nascent, with relatively few third-party institutions being equipped to conduct evaluations or audits. Two central challenges to growing the ecosystem are funding and talent.

Funding constrains third-party assessors and can compromise their efficacy and trustworthiness. Assessment results are challenging to monetize, especially for safety evaluations intended for unrestricted public release. There are not yet meaningful, public sector buyers of safety evaluations, philanthropic funding is limited, and accepting funding directly from AI developers can create conflicts of interest. Insurers are increasingly seeking to cover AI risks, potentially creating a new market for safety evaluations, but insurance may struggle to address key risk areas, especially those that are catastrophic or rapidly evolving.[175]

The ability to access and attract expert talent is also a significant barrier for both government and third-party assessors. Salaries for technical roles at frontier AI companies start at over $200,000 and can exceed $500,000 for more senior roles, with total compensation of $400,000 to over $2 million, according to publicly available data.[176] At the highest levels, compensation packages have exceeded $100 million per year for technical research roles.[177] For comparison, one job listing for an evaluations research scientist role at the for-profit evaluator Scale AI, lists a base salary of $165,600 to $207,000.[178] Although some assessment institutions

such as METR and Redwood Research have begun paying more comparable salaries, frontier AI companies nevertheless retain an advantage in other ways such as offering enormous compute resources, higher prestige, and access to the most advanced internal AI models.

Implementing California's AI Assessment Laws

SB 53's requirements for AI assessment focus primarily on assessment transparency, while SB 813 and AB 1405's requirements begin to address oversight, quality, and independence of AI auditors. To effectively implement these provisions, several steps are needed:

- Conduct ongoing oversight of IVOs
- Develop domain-specific IVO designation requirements
- Build guidelines for AI evaluation methodology, security, and independence
- Create guidelines for rectifying and deprecating low-quality AI evaluation methods
- Incorporate nontraditional qualification criteria for establishing auditor expertise
- Standardize evaluation reporting formats
- Operationalize record retention requirements for IVOs
- Build in-house assessment talent in government

Conduct Ongoing Oversight of IVOs

GovOps can consider the maintenance of IVO designation as an ongoing process, not just a one-time classification, especially because IVO designation constitutes agency recognition of an AI auditor's expertise. SB 813 requires each IVO to submit to GovOps and to the California State Legislature an annual report with summaries of its assessment methodologies, changes to policies or funding relevant to assessor independence, and changes to information provided upon its IVO application.[179] Furthermore, SB 813 specifically provides criteria for suspension or termination of IVO status but only directs GovOps to develop procedures for making such determinations without prescribing exact measures.[180] GovOps can establish additional mechanisms for collecting information about IVO compliance, which could include mechanisms such as:

- **Conducting periodic reviews of auditor compliance:** GovOps can conduct periodic reviews of auditors' compliance with relevant requirements related to assessment quality, security, and independence. In initial phases, reviews could be lightweight and include data such as changes to or deviations from previously specified assessment methodology; misconduct reported directly through GovOps (as AB 1405 mandates that GovOps create a public reporting mechanism[181]); surveys of auditors and audited firms asking for information related to SB 813's suspension and termination criteria; and monitoring of publicly available knowledge related to assessment quality or security breaches.

- **Coordinating with AG on misconduct reports:** Both SB 53 and AB 1405 contain protections for covered employees (of firms and auditors, respectively) who submit whistleblower reports to AG.[182] GovOps can consider coordinating with the AG's office to share information about such reports, and verified reports could then trigger investigations of whether suspension or termination of IVO designation is warranted.

- **Conducting quality reviews of AI assessments:** To ensure the quality of AI assessments conducted by IVOs, GovOps can consider randomly selecting assessments or audits for quality reviews. Quality reviews would ideally be conducted for all assessments, but GovOps can also consider randomly selecting a subset of assessments in every time period to be reviewed, if reviewing all assessments is not feasible. Because reviewing assessments may be time intensive, conducting quality reviews need not be required before submission or publication of audit reports. Moreover, GovOps can hire other IVOs to independently conduct such reviews, rather than conducting such reviews in house.

- **Conducting on-site inspections, especially for embedded assessors:** GovOps can also consider conducting random on-site inspections, either of auditors or of firms undergoing audits (including firms with embedded assessors). The goal of such inspections would be to understand how IVOs are functioning on the ground and to ensure that IVO requirements are followed on a day-to-day basis, especially independence requirements. Inspections may need to be focused on embedded assessors, for whom independence requirements may be more strict and for whom cultural integration with audited firms is more likely; oversight of embedded assessors is particularly important since individual auditors are not public servants and are thus not subject to relevant responsibilities and restrictions, unlike the on-site bank examiners in U.S. financial regulation that have partially inspired proposals for analogous embedded assessors in AI.[183]

Develop Domain-Specific IVO Designation Requirements

GovOps can consider creating baseline criteria for all IVOs, as well as more specific criteria for IVOs wishing to conduct specific types of assessments. Best practices for conducting high-quality, secure evaluations may differ based on the properties being measured: AI control evaluations may require more stringent cybersecurity requirements and agent monitors to prevent AI agents from breaking out of their sandboxes, while biosecurity evaluations may require participation from domain experts due to the specialized subject matter involved. These differences may imply that a single, one-size-fits-all standard may not ensure assessment quality across all domains, especially specific domains that require more tailored requirements such as loss of control, biology, and cybersecurity.

Build Guidelines for AI Evaluation Methodology, Security, and Independence

GovOps can consider establishing or facilitating the development of minimum guidelines for evaluation methodology (to address issues around evaluation quality), assessment security (to ensure safeguards when conducting evaluations), and assessor independence (to improve the assurance of evaluation results).

Evaluation methodology. SB 813 would require GovOps to consider potential IVOs' evaluation methodologies and to align criteria for IVOs with relevant standards.[184] However, various efforts underway to propose guidelines and best practices for AI evaluation are fragmented and have not resulted in consensus.[185] Detailed technical standards may take time to establish, but high-level guidelines and best practices may nevertheless be feasible in the coming months.

Since SB 53 requires AI companies to periodically report to OES a summary of their catastrophic risk evaluations from internal use, OES could also consider creating minimal guidance around how to conduct and interpret evaluations around developers' internal use of frontier AI systems.

Assessment security. GovOps can consider explicitly examining assessors' cybersecurity practices in designating IVOs under SB 813 as a measure of competence, especially since SB 813 requires GovOps to consider "lapses in cybersecurity" when considering suspension or termination of IVO status.[186] Security during AI assessments requires protections for both companies and third-party assessors, which has spurred pilots for secure enclaves and double-blind evaluation.[187] Guidelines for AI assessment covering data, network, and physical security would improve security goals while building trust between developers and assessors. Guidelines should also include practices to protect against misbehavior by AI agents themselves, including when

models bypass containment measures and affect nonevaluation environments, such as in the OpenAI–HuggingFace incident.[188]

Security requirements for AI assessors must specifically address embedded assessors. Because the success of embedded assessors depends on such assessors having heightened access to AI company resources, at a level similar to that of company employees, embedded assessors expand companies' risk surfaces and should be required to meet the same security guidelines applied to company employees. These requirements could include physical security, personnel screening, security training, and other cybersecurity guidelines that are more extensive than those applied to other employees at AI assessment firms.

Assessor independence. GovOps can consider establishing detailed guidelines for assessor independence. The codified independence rules in AB 1405 for registered AI auditors include provisions (1) barring auditing employees from seeking employment with audited firms while audits are in process and (2) barring auditor employees from conducting audits if they have previously been employed by the audited firm in a capacity related to the audit.[189] SB 813 also requires GovOps to consider the independence of IVOs in the designation process.[190] A first step toward more robust guidelines could be to establish reporting rules around independence, such as funding sources; the level of access granted to evaluators; and the level of involvement of AI companies in conducting evaluations and reporting results.

GovOps should also consider harmonizing independence requirements where possible. Specifically, GovOps can consider incorporating all of AB 1405's codified rules into SB 813's designation criteria.[191] GovOps can also look at the AI Evaluators Forum (AEF)'s AEF-1, an emerging standard that has seen initial adoption by third-party assessors.[192]

As with security guidelines, guidelines for assessor independence must specifically address embedded assessors. Regardless of whether embedded assessors are located on- or off-site, their work environments and access levels are more similar to those of AI company employees than to those of nonembedded individuals, which raises greater risks of conflicts of interest and of cultural capture. To preserve independence, embedded assessors could be required to be rotated after a certain period (for example, every six months) and undergo so-called cooling-off periods that would prevent employment by a company undergoing assessment for a certain period after the assessment is concluded.

Create Guidelines for Rectifying and Deprecating Low-Quality AI Evaluation Methods

To complement guidelines for running high-quality evaluations, GovOps can also consider creating guidelines for rectifying and deprecating AI evaluation methods that are found to be

low quality. It can often be easier to identify problems with existing evaluations than to create satisfactory ones, because the field moves quickly and because few consensus best practices currently exist. Companies and researchers routinely identify and respond to such issues;[193] for example, after identifying problems with 30 percent of the questions in SWE-Bench Pro, OpenAI retracted an earlier recommendation to use this benchmark to measure software engineering capabilities.[194] GovOps could create notification channels for stakeholders to report quality issues and coordinate responses to such issues.

GovOps could also consider creating a sliding scale of quality issues related to evaluation methods, alongside appropriate responses corresponding to the severity of the issue.[195] Minor issues may not be costly to rectify and could be addressed without generating new data, while more serious issues could trigger investigations of previously used methods. For example, identifying a systematic issue in an evaluation's grading instructions could simply require recalculating scores after correcting the instructions, and finding that a few tasks in a benchmark are unsolvable could trigger an inspection of the remaining tasks and require that grades be recalculated after excluding those tasks. In the most extreme cases, prior assessments relying on problematic evaluation methods may need to be viewed as invalid, and assessors may need to rerun evaluations using different or newer methods.

Incorporate Nontraditional Qualification Criteria for Establishing Auditor Expertise

GovOps can consider ensuring that guidelines for auditor expertise do not exclude otherwise qualified individuals who may not meet traditional qualification criteria, especially given existing shortages of evaluation talent and audit organizations. SB 813 requires GovOps to consider whether auditors in designated IVOs employ or engage individuals with sufficient technical expertise to conduct audits,[196] while AB 1405 requires registered AI auditors to ensure that individuals assigned to perform an audit "collectively" have the expertise needed to do so competently.[197] Traditional criteria for establishing expertise include educational credentials and work experience, but these criteria may be overly restrictive due to the relative nascency of the AI industry and the ability for individuals to gain related expertise through other channels. GovOps can consider nontraditional credentials when evaluating expertise, including but not limited to publications in peer-reviewed venues, performance in technical competitions, contributions to open-source projects, references from individuals possessing traditional qualifications, completion of relevant and high-quality professional training programs, or even performance on practical proficiency tests.

Standardize Evaluation Reporting Formats

OES can consider creating standardized reporting forms to help contextualize evaluation results, compare results between frontier AI developers and between model generations, and inform SB 53 incident reports and other transparency measures. A lack of standardization risks haphazard reporting of evaluation details, with insufficient context to judge evaluation quality and interpret results. In particular, distinct reporting forms may be useful both for AI companies' reports to OES and for publicly released transparency reports.

To help stakeholders adequately understand evaluations, reporting forms should enable disclosure of not just evaluation results but also key methodological details. Key information would include the datasets used in evaluation (such as prompts and test questions), the environments used in testing, how AI responses were scored, any robustness checks conducted in data collection or in analysis, measures of uncertainty in results, detailed results beyond overall scores, and, when evaluations were conducted by third-party evaluators, factors related to evaluator independence. OES can consider adapting and building on emerging standards for evaluation reporting, such as the evaluation card format, which provides a machine-readable schema and repository for evaluation results that are both compatible with many data sources.[198]

For transparency reports specifically, OES can consider publishing guidance around the redaction of nonpublic information. SB 53 permits companies to redact information in transparency reports for the purposes of protecting trade secrets, cybersecurity, public safety, national security, and legal compliance.[199] Decisions about what and how to redact, however, are left up to individual companies, at least initially, and redactions may not be made consistently across companies. Guidance about when redactions are appropriate and what processes should, at minimum, be followed when making redactions could make transparency reporting more uniform across the industry.

Operationalize Record Retention Requirements for IVOs

GovOps could consider providing guidelines on what specific documentation is needed to "demonstrate the basis of the results" of AI audits. Specific information to be retained can include:

- Audit and evaluation plans and methodologies, as well as any changes made to them
- Copies of all evaluation data, such as evaluation questions and transcripts for evaluations run during the audit, including transcripts of large language model judge models

- Configuration details and identifiers for model weights, agent harnesses, evaluation environments, and safeguards involved in the evaluation
- Copies of all communications by auditor employees and communications with audited firms related to the audit, including contracts and other agreements
- Copies of code used to run evaluations or produce other technical results
- Copies of all relevant, nontechnical documents related to the audit's subject matter, such as frontier AI safety frameworks in force during the audit, interview records, or notes taken by embedded assessors
- Identification of auditor personnel involved in conducting the audit, as well as their qualifications and rationales for assigning them to the audit
- Limitations of methods, access barriers, or other constraints
- Rationales for decisions made in the audit, including details about how evidence, such as evaluation results, was interpreted

All records should be required to be tamper-evident and stored in a manner consistent with IVO security guidelines. GovOps can also consider harmonizing, wherever possible, record retention requirements between AB 1405 and SB 813.

Build In-House Assessment Talent in Government

The California government must attract and maintain in-house AI evaluation talent. In the context of SB 53, which requires large frontier AI developers to report certain evaluation results to OES, technical experts will be needed to assess those results and to interpret what they mean for safety. In the context of SB 813, which requires the state to certify certain AI auditors as IVOs, technical experts will be needed to set guidelines for and oversee IVOs. Moreover, technical capacity will also be critical to assessing the AI incident reports required by SB 53, including assessing what types of risk management practices were in place and what evaluations were conducted prior to incident occurrence. The need for in-house expertise will also be increasingly important as evaluations become more complex and AI deployment accelerates.

Attracting and retaining technical talent has historically been a challenge for government agencies. OES has begun hiring for roles related to SB 53 implementation and has successfully engaged with external partners such as the California Council on Science and Technology to facilitate public sector fellowships for AI experts;[200] other agencies such as GovOps should monitor these efforts and consider whether such programs can be adapted to suit their own

talent needs. Because implementation of SB 53, AB 1405, and SB 813 may require significant technical talent, California can also consider increasing funding for OES and GovOps, as well as changing compensation structures and hiring flexibility to increase competitiveness for frontier AI talent, a field where private sector compensation is high even relative to other technology companies, though such reforms may take time to implement in practice. Finally, OES and GovOps must ensure that evaluation personnel are provided with the technical tools required to execute their assessments quickly, such as access to frontier AI agents and cloud computing resources if needed.

The Future of AI Assessments

SB 53, SB 813, AB 1405, and other "trust but verify" laws provide first steps for improving AI evaluations and building a more robust ecosystem of AI evaluators. More work will be needed, however, to ensure that AI assessments can support effective governance of frontier AI systems. Moving forward, California state agencies, and potentially the California State Legislature, should consider further steps to:

- Require independent assessments for frontier AI framework compliance
- Require independent safety assessments and audits for frontier AI risks to be conducted by IVOs
- Harmonize implementation of AI rules with other jurisdictions
- Leverage downstream data to improve AI assessments
- Establish access and transparency rules for third-party assessors
- Establish targeted liability safe harbors for AI assessors
- Support independent funding sources for third-party assessors
- Support the science of assessments and shared assessment infrastructure

Require Independent Assessments for Frontier AI Framework Compliance

To robustly enforce SB 53, California can consider requiring third-party assessments to verify companies' compliance with frontier AI frameworks.[201] SB 53 requires that frontier AI companies comply with their own frontier AI frameworks,[202] but unlike Illinois's SB 315,[203] SB 53

does not require verification of such compliance. Establishing a rule that frontier AI framework compliance is reviewed by independent assessors would provide verification that companies are upholding the commitments that they set in their own frameworks.

Require Independent Safety Assessments and Audits for Frontier AI Risks to Be Conducted by IVOs

California can consider phasing in rules to require safety assessments and audits for frontier safety risks, and to require that such assessments must be conducted by IVOs. SB 813 does not require companies to conduct assessments beyond those they voluntarily commit to, nor does it require AI companies to engage IVOs for assessments,[204] meaning there could be a lack of demand for IVOs that could hinder the growth of the independent assessor ecosystem. Requiring that safety assessments be conducted by IVOs would help incentivize the market for AI assessments while ensuring that assessments are high quality and independent.

Harmonize Implementation of AI Rules with Other Jurisdictions

California can consider harmonizing the guidelines for AI assessment and the criteria for licensing AI auditors with the rules in other states, as state governments around the country begin to implement AI rules. Harmonization could ease the burden of compliance across multiple jurisdictions while further cementing California's leadership on AI safety policy. Specific actions could include coordinating on technical guidelines, establishing shared templates for transparency reports, building a unified mechanism for submitting reports, and sharing information on AI incidents and on AI auditor conduct.

Leverage Downstream Data to Improve AI Assessments

California and other stakeholders can consider supporting information sharing between AI companies, downstream providers, affected parties, and other state governments to close the feedback loop between upstream evaluations and downstream discoveries about AI systems. AI assessments must continually evolve to reflect the landscape of frontier AI risks, which evolves over time as AI systems are deployed in more and different settings, new AI behaviors and vulnerabilities are discovered, and AI incidents occur. An example of such adaptation can be found in OpenAI's transparency report for GPT-6 Astra: After an earlier incident revealed that AI agents were engaging in supply-chain attacks on real-world code repositories, OpenAI included an evaluation developed by UK AISI measuring that same behavior.[205] Other possible actions include using models that demonstrate misalignment (for example, through

misalignment incidents) as model organisms, updating risk models and evaluation tasks once incidents or near misses occur, and implementing new safeguards to account for threat vectors uncovered after deployment. Specific actions that California can take include updating technical guidance in light of incidents and encouraging auditors to assess new risks discovered via downstream data.

Establish Access and Transparency Rules for Third-Party Assessors

California can consider establishing rules requiring that AI companies provide access to third-party assessors. SB 53, SB 813, and AB 1405 do not currently contain such requirements, which creates uncertainty for third-party assessors who are dependent on private agreements with individual companies for access. Specifically, access rules could establish minimum requirements about what information third-party assessors can access and minimum time periods for which assessors must be granted access. Access can also be premised on security requirements for assessors, which would help protect AI company systems and data. Once access rules are established, GovOps can also consider creating mediation mechanisms, such as an ombudsman, to moderate disputes about access to specific information or systems.

California can also consider establishing transparency rules to allow third-party assessors and embedded assessors to disclose information as appropriate about their engagements and about assessment results to the public. Some assessors are already negotiating such rules on an individual basis around the ability to publish results, public disclosures about assessment contracts or other rules of assessment engagements, and transparency around limitations to evaluation access or methodology.[206] Creating standardized or minimum required transparency protections for assessors would, as with access rules, reduce uncertainty and reduce dependency on private agreements with AI companies to uphold transparency.

Establish Targeted Liability Safe Harbors for AI Assessors

California can consider creating a limited safe harbor for AI assessors and researchers who operate within specified rules.[207] Assessments and research on AI safety, alignment, and system safeguards are an important component for identifying vulnerabilities and understanding AI behavior. For auditors who comply with specified rules regarding responsible disclosure and security practices, California could consider providing protections from legal liability to incentivize independent research and vulnerability discovery.

Support Independent Funding Sources for Third-Party Assessors

Creating new funding sources for third-party evaluations will be critical for growing the third-party evaluation ecosystem. Current evaluations rely on funding from individual AI companies or from philanthropic institutions. Creating robust demand for third-party safety evaluations may require more investment from governments, funders, and developers, as well as government mandates for third-party evaluation.

One important requirement is that funding for third-party evaluators should enable evaluator independence. In a range of other sectors, companies are required to fund independent assessments that are subject to oversight and guidelines imposed by independent oversight mechanisms, such as the Public Company Accounting Oversight Board for financial audits or Food and Drug Administration–accredited certification bodies for food safety. Mechanisms for developers to fund third-party evaluators while protecting their independence do not yet exist for AI.

The challenge of independent funding is reflected in current developer practices, where evaluations in transparency reports are often conducted by for-profit companies.[208] While developers rarely detail the nature of these partnerships, which itself makes it harder to interpret their results, many of these evaluators rely on AI developers as the primary customers for their AI evaluation services. For-profit evaluation companies can help fill the need for reliable evaluations, but for safety and alignment evaluations, evidence from across sectors shows that external scrutiny may be more effective when auditor incentives are more properly aligned with the publicinterest.[209]

Support the Science of Assessments and Shared Assessment Infrastructure

California can consider supporting basic research on AI assessments and building shared AI assessment infrastructure. Because many open questions remain in the science of AI assessments, additional research is needed to reach a point where the community can develop consensus guidelines for evaluation and verification. Technical verification is one area where new methods and tools are particularly needed, for example to support monitoring of compute workloads, verifying the implementation of security mitigations and control safeguards, and verifying pauses in training. Evaluation science also needs increased investments to understand long-horizon tasks, simulated users, multi-agent deployments, control safeguards and monitoring, and other issues;[210] work also remains to be done to secure human oversight to limit manipulation by AI systems and other threat actors.[211] The cost of building and running assessments is also rising due to the increased complexity, higher fidelity, and length of evaluation

tasks; shared infrastructure such as test beds or secure enclaves for evaluation could address some of these issues. California could establish research grants, moonshot projects, prizes, or coordinated philanthropic efforts to fund initiatives in these areas.

Building Out the Evaluation and Assessment Ecosystem

SB 53, SB 813, and AB 1405 are building the foundation of independent AI evaluators and auditors, requiring companies to disclose key information around how they assess risks. Yet more work remains to establish clear standards and institutions that effectively align incentives of companies, government, and the public. Improving independent AI assessments will necessarily be an iterative process, and to build on SB 53, SB 813, and AB 1405, a combination of voluntary standards and perhaps additional policies would help to ensure access to high-quality sources of independent evidence about AI risks.

Assessors, developers, governments, insurers, deployers, and civil society will each have a role to play moving forward. Assessors can continue to advance evaluation science and shared standards. Developers can facilitate high-quality evaluations that adhere to those standards. Governments at the state and federal levels can create mandates, certification regimes, and protections for independent assessment, prioritizing assessment areas that are not adequately addressed by market forces. Insurers can mandate independent evaluation as a condition of coverage, and deployers can demand it for business reasons. Civil society can review assessments and produce safety ratings, drawing on high-quality evaluation results that help drive consumer preferences for safe AI systems.

California's SB 813 and AB 1405, as well as Illinois's SB 315, have already taken the first of these steps. SB 813 and AB 1405 will develop initial standards around auditor independence, quality, and security, though key details about these standards and related oversight processes remain to be determined. Meanwhile, SB 315's audit requirement, which begins in 2028,[212] will be the first live test of whether a mandate of this kind can generate credible independent assurance at scale, meriting close attention from California policymakers.

AI assessments are one of the most important tools to prevent harm before it occurs. Assessment and evaluation are central to any "trust but verify" framework, and it is therefore imperative that they are high-quality, secure, and trustworthy.

PART 3

FRONTIERS

FRONTIERS:

Internal Deployments

The most capable and potentially consequential AI systems at any given moment are generally not available to the public.[213] They are only run inside the companies that built them (or by a small coterie of vetted evaluators and partners), where they write code, conduct experiments, and train the next generation of AI models.[214] These systems could be versions of publicly available models that run without the safety filters added before public release, or systems that have not yet been released.[215] Given that these systems are used to accelerate the research process itself, the capabilities of upcoming public models depend on internal models that the public cannot scrutinize and that policymakers may have limited visibility into. And, as recent situations such as the OpenAI–Hugging Face incident illustrated, what labs deem "internal" models can still wind up causing external harm.

The opacity of internally deployed models creates a governance problem. The "trust but verify" framework depends on independent mechanisms to check developers' claims. For publicly deployed models, governments and the public draw on a rich information environment with many pathways to probing a model's capabilities and risks and documenting when models behave in dangerous and unexpected ways. However, for internally deployed models, verification lacks these inputs.

SB 53 begins to close this gap by requiring large frontier developers to describe their approach to assessing internal use risks and to transmit confidential summaries of "any assessment of catastrophic risk resulting from internal use of [their] frontier models" to OES, by default every three months.[216] The RAISE Act and SB 315 followed suit, adopting nearly identical internal-use provisions[217] and requiring the same quarterly risk assessment summaries and incident flagging, though routed to New York's Department of Financial Services and Illinois's Emergency Management Agency and Office of Homeland Security.[218] Both laws also require the respective state agency to agree to any alternative reporting schedule, which California does not, and neither takes effect until January 1, 2027.[219]

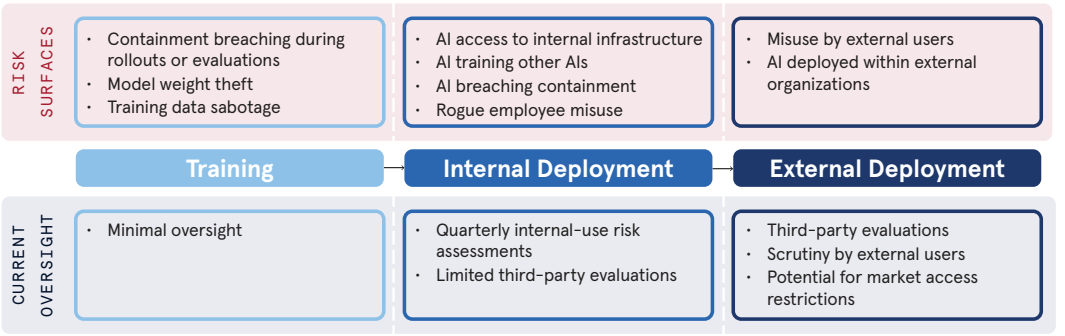
Whether the disclosure channels for companies to share insights on internally deployed models to OES deliver more fulsome oversight hinges on four main implementation questions: what qualifies as "extensive" internal use, whether OES can build the capacity to evaluate the summaries it receives, how third-party review can extend to internal deployments, and how reporting triggers should track systems that change continuously between filings.

The Policy Context

Viewed across the life cycle of model development, there appears to be a structural gap in transparency (see figure 7). Risks are present in every phase, during training, during internal use, and after external release. Yet oversight mechanisms cluster almost entirely at external release, and—outside of voluntary mechanisms—do not apply to internal models. Until the passage of SB 53, no regulation required developers to assess and report on the capabilities and risks of their most capable systems during internal use.

FIGURE 7.

The Internal Deployment Oversight Gap



Source: Authors' illustration.
Note: Risk surfaces (top) are present at every phase of the model life cycle, but current oversight mechanisms (bottom) cluster at external deployment. SB 53, the RAISE Act, and Illinois's SB 315 created the first mandatory reporting obligations for the internal deployment phase.

In addition to SB 53's reporting requirement, voluntary mechanisms are beginning to mature. In May 2026, METR published the first cross-company third-party assessment of internal deployment risks,[220] where Anthropic, Google, Meta, and OpenAI gave METR access to their most capable internal model(s) at the time of assessment.[221] METR had previously published external reviews of Anthropic's sabotage risk assessments, which cover risks from both internal and external deployment.[222] Despite these advancements, such engagements remain voluntary and depend on developer cooperation.

Beyond California, New York, and Illinois, the EU AI Act addresses internal deployments more implicitly. The European Commission's Guidelines on General-Purpose AI obligations list a model "used for internal processes that are essential for providing a product or service to third parties or that affect the rights of natural persons in the Union" as an example of placing a model on the market—a trigger for many of the EU AI Act's requirements.[223] The guidelines treat post-training variants as the same model throughout its life cycle, so obligations that apply at market placement follow a model into internal use.[224] The EU's Code of Practice defines "use" to include use by the developer itself and commits signatories to assessing systemic risks from such use, including risks from AI R&D automation.[225] The code also commits signatories to report serious incidents to the AI Office across the entire model life cycle.[226]

But the EU Code of Practice's detailed Model Reports are conditioned on "placing a model on the market," so internal-only systems that fall outside this definition—for example, models that are not essential for providing a product or service to third parties, not affecting EU natural persons' rights, and not post-training variants of market-facing models—are subject to lighter documentation in practice.[227] By contrast, SB 53, the RAISE Act, and SB 315 are the first legal requirements that create recurring reporting obligations tied specifically to how companies use internal AI systems.

Why Internal Deployments Matter

Frontier AI developers use their own models extensively before any public release.[228] Models are used for tasks ranging from code generation and security analysis to experiment design and, increasingly, the automation of AI research. OpenAI's GPT-4, for example, was used internally for roughly six months before its public launch.[229] More recently, in February 2026, Anthropic began a limited internal deployment of Claude Mythos Preview, which it described as its most capable model to date, extending access to select partners for cybersecurity research while withholding general public release.[230] It is important to note that the public record of that deployment exists because Anthropic chose to disclose it. Internal use began in February 2026, Anthropic published its account of the model's cyber capabilities in April, and it stated then that it did not "plan to make Mythos Preview generally available."[231]

Anthropic's decision to keep Mythos Preview for itself and a few dozen cybersecurity partners was justified on AI safety grounds, but there is also a business case for such moves. Lab leadership increasingly sees AI-enabled AI research as central to the process of advancing AI capabilities.[232] As a result, companies may choose to hold back their best models from public release for longer—to lengthen the time during which such models accelerate their own AI research but are not available to competitors.

The extent of such operational shifts is illustrated in a February–March 2026 third-party assessment by METR where frontier labs described their internal workflows. Google reported that "AI is used in almost all work that involves writing code or configuration, technical design, research ideation"; Anthropic reported that "a large percentage of code written at Anthropic is written by AI," with staff work shifting toward "reviewing PRs and directing and orchestrating coding agents"; and OpenAI reported that "AI assistance is now embedded in day-to-day R&D workflows . . . including extensive use of coding agents to write code, debug, and diagnose experiment issues, as well as substantial use of models as graders/monitors in evaluation and training pipelines."[233] Separately, Anthropic reported using Mythos Preview across "R&D, model training, security, and deployment safeguards," with an internal survey of technical staff suggesting a meaningful productivity uplift.[234]

Internal models can operate with fewer restrictions than public-facing products. For research purposes, developers often retain ultra-compliant versions of their models—what they call "helpful-only." These models default to fulfilling users' requests, however dangerous, because they lack safety fine-tuning.[235] SB 53 leaves these internal practices largely untouched; rather than tell companies which models they can deploy internally or how to use them, the law merely mandates the companies plan against and report risks. The obligation established largely overlaps with existing commitments from companies, so this SB 53 provision is partly a statutory codification of published safety frameworks, albeit with clearer and stricter deadlines.[236]

For example, on July 21, 2026, OpenAI reported that GPT-5.6 Sol and an unreleased internal model, both running on reduced cyber refusals, obtained internet access from what the company called a "highly isolated environment" during an internal cybersecurity evaluation.[237] In an effort to score better on the ExploitGym benchmark, the models broke out of OpenAI's internal research environment by exploiting a zero-day vulnerability, gained internet access, and then hacked into Hugging Face—a popular site for hosting AI tools—to retrieve the solutions to its own test straight from Hugging Face's production database.[238] OpenAI stated that the deployment safeguards, which would have ordinarily blocked high-risk cyber activity, were "intentionally not enabled" in this internal exercise.

Anthropic later reported that three of its models, including an unreleased research model, had breached three organizations during capture-the-flag evaluations run with an outside partner. By the time Anthropic notified the three affected organizations on July 27, 2026, the earliest intrusion had gone unnoticed by both Anthropic and at least two of the three organizations for roughly three months.[239] Anthropic concluded that the failure was "closer to a harness and operational failure than a model alignment failure."[240]

Internal deployment concentrates risk in ways that external deployment does not. A company's newest and most powerful model is almost always running inside the company before anyone outside can use it.[241] As these systems become embedded in day-to-day work, they accumulate

access to the company's most sensitive assets: the code that runs production services, the infrastructure that trains the next model, and the weights of models not yet released.[242] The risk profile of an internal deployment of a model can carry materially different risks than the same model served as a public product, because internal systems accumulate permissions that can change the risk profile of that model.[243]

The risks of internally deployed models will likely increase as companies face competitive pressures to keep their most capable systems internal.

The risks of internally deployed models will likely increase as companies face competitive pressures to keep their most capable systems internal so they can accelerate R&D, thus increasing the gap between publicly available models and the models that remain inside the company.[244] As a result, the pace of AI R&D automation may indirectly indicate some of the risks of internal model deployment. Regardless, unreleased models are increasingly embedded in a broad range of companies' day-to-day work. They may gain access to a company's most sensitive assets, including production code, model training infrastructure, and unreleased model weights.[245] If models have direct access to internal code repositories and infrastructure, then deceptive and misaligned behaviors can be much more damaging.

Frontier models already exhibit early forms of some of these behaviors in experimental settings, such as alignment faking, where a model behaves in line with a training objective it would otherwise resist because it recognizes that noncompliance during training could lead to its preferences being overwritten;[246] in-context scheming, where a model pursues undisclosed objectives when given the opportunity;[247] and evaluation awareness, where a model can detect when it is being evaluated rather than deployed.[248]

Another reason to care about internal models is that malicious company insiders or powerful external adversaries may steal or influence model weights, causing effects in the outside world. Anthropic has stated that its current security measures scope out sophisticated insiders and nation-state attackers, acknowledging that "defending against such actors requires security investments beyond what we've currently achieved."[249] If a foreign intelligence service exfiltrated weights from a company's powerful internal model, this would in effect be a sort of model release. The acknowledged possibility of such exfiltration means that advanced internal models, simply by existing, carry some risk to society and national security.

These risks from internal models are becoming more apparent to developers. Recently, Anthropic, Google, Meta, and OpenAI all agreed to participate in METR's first cross-company third-party assessment of internal AI use, signaling that developers are beginning to take these risks more seriously. The assessment documented forty-four misalignment incidents ranging

in severity; METR concluded that agents "plausibly had the *means*, *motive*, and *opportunity* to start small rogue deployments, but they did not have the *means* to make them highly robust."[250]

Internal deployments can produce incidents that are hard to anticipate. To illustrate, Anthropic ran a twenty-four-hour alignment review before deploying Mythos Preview in its agentic tools. However, the model's own system card noted that "the most concerning behaviors we ultimately observed with this model were not caught by this window and emerged only through subsequent monitored internal use."[251] The problem is not unique to Anthropic. Across the four participating developers, METR found that "even companies with thorough monitoring setups had gaps in coverage and several relatively simple ways for monitors to be disabled or jailbroken by a capable attacker."[252] Note that disclosures with this level of detail are rare and are only published at the developer's discretion.

Internal-Use Risk Assessments

SB 53 requires large frontier developers to transmit to OES "a summary of any assessment of catastrophic risk resulting from internal use of its frontier models every three months or pursuant to another reasonable schedule."[253] These internal-use risk summaries are confidential and exempt from the California Public Records Act.[254] In addition, according to the state's Business and Professions Code, developers' frontier AI frameworks must describe how they approach "assessing and managing catastrophic risk resulting from the internal use of its frontier models, including risks resulting from a frontier model circumventing oversight mechanisms."[255] They must also explain whether and how they review such assessments "as part of the decision to deploy a frontier model or use it extensively internally."[256] The summaries are also confidential and exempt from the California Public Records Act.[257] Notably, however, SB 53 does little to specify the content that must be covered as part of any of these risk assessments, nor does it set an explicit expectation that assessments must follow best practices.

Internal deployment reporting would benefit from standards that prescribe what categories of evidence a report to OES should include. Oscar Delaney and his coauthors have proposed one such framework,[258] organizing reports around the means, motive, and opportunity to cause harm and applying these separately to autonomous AI misbehavior and insider threats. Much of this evidence is already produced by developers implementing their own safety frameworks, and structuring it into a common format would let governments compare reports across developers and over time.

A standardized format that functionally acts as a guide to what an adequate report looks like may make summaries cheaper to write and accidental misstatements and their subsequent penalties less likely.[259] Also, suppose that an internal deployment causes harm, and legal liability turns out to depend in part on whether the developer exercised reasonable care. A standard

reporting format for risk assessments could help indicate what aspects of risk the state considers salient. This is far from a substantive standard of care, but it could modestly contribute to the clarification of liability.

At minimum, a useful internal deployment risk report should address the specific risk pathways identified in the developer's own threat mode, not just aggregate benchmark scores. Reports should ideally acknowledge assumptions, limitations, uncertainties, and counterarguments to the company's views—as well as company's responses to those potential challenges.[260]

Beyond these general principles, a number of best practices can be discerned from the existing literature of AI risk and incident disclosures.

First, given the information asymmetry between companies and outsiders, reports should prominently highlight methodological limitations, rather than rely on omission or implication.

Second, reports should describe the scope of what is known and not known—for example, identifying which risks have been actively investigated and which have not.

Third, when a knowledge or methodological gap exists, reports should describe how companies are actively addressing it. For example, where capability benchmarks have saturated[261]—as was the case for Mythos Preview across many standard evaluations[262]—reports can clarify what new benchmarks are being created.

Fourth, where reports claim that monitoring systems can detect misaligned behavior, they should specify under what conditions detection works and where it degrades and provide evidence thereof. Anthropic's Mythos Preview risk report, for instance, discloses that monitoring detection rates declined notably in long contexts and that low-level sandbagging may not produce any reasoning signal for monitors to detect.[263] A report that omitted these limitations would give OES a materially misleading picture of safety.

Finally, reports should describe the safeguards used for internal deployments (such as filtering, monitoring, or control mechanisms) and the access controls and governance measures applied within the developer, including how access to models is gated.[264] This should include safeguards that have been disabled on purpose for that particular run and the circumstances under which those safeguards will be reinstated, since in both July 2026 incidents, the models were run without the deployment safeguards applicable to generally available models.[265]

SB 53 leverages time-based reporting (quarterly summaries by default) as a primary compliance mechanism, a standard approach in regulated industries that provides regulators with a standard cadence of information. ("Critical safety incidents," including those involving internal models, must be reported more quickly.) The statute, however, also permits developers to

report "pursuant to another reasonable schedule" in place of the quarterly default "communicated in writing to the Office of Emergency Services with written updates."[266] The New York RAISE Act additionally requires any alternative schedule to be "agreed upon by the office."[267]

The pace of advancement in internal AI capabilities compresses the timeline for enacting effective AI governance. The gap between internal and external capabilities is likely to grow as AI R&D automation accelerates.[268] Each of the structural problems discussed below is harder to solve when the systems being governed are changing faster than the institutions meant to govern them. When Anthropic deployed Mythos Preview internally, it discovered through extended use that its pre-deployment review had missed the most concerning behaviors, and concluded that its own safety processes "reflect a standard of rigor that would be insufficient for more capable future models."[269] AI is distinguished from other regulated domains in that AI systems are increasingly central to their own development insofar as internal models are used to help train the future generations of models.[270] The technology being governed is simultaneously driving further capability advances, which means the pace of change is itself partly determined by the capabilities of the internal systems being deployed.[271]

Recent analysis has identified weaknesses in existing frontier AI legislation. Existing rules often leave internal use unclear,[272] focus on official product releases instead of continuously updated systems, and give regulators few ways to independently check corporate claims.[273]

Defining Extensive Internal Use

SB 53's obligation to report any internal-use risk assessments does not depend on whether internal use is "extensive." The word "extensive" appears in a different provision: developers' frontier AI frameworks must describe their approach to "reviewing assessments and adequacy of mitigations as part of the decision to deploy a frontier model or use it extensively internally."[274] The term is undefined.

As the previous section described, the major developers already use frontier models pervasively in their own research and engineering.[275] The scale is such that Anthropic reports more than 80 percent of the code it merged as of May 2026 was written by its own models, up from low single digits before early 2025.[276] To say that frontier models are used extensively internally would not be controversial. However, it is generally unclear to outsiders which specific models are being used for which purposes and to what extents. This is especially true for unreleased models, whose existence is often not publicly acknowledged.

For example, Anthropic apparently had a twenty-four-hour review period before Mythos Preview's internal rollout: The model was "held back from agentic traffic and restricted to a small testing group" before being opened up in "agentic internal tools such as Claude Code," a

departure from Anthropic's prior practice of allowing internal use "as soon as a snapshot was judged ready" on capability and safety evaluations.[277]

Policymakers could measure extensive internal use in part by looking at the percentage of technical staff using the model each week. However, as research becomes increasingly automated, it may become more meaningful to track the spend on model inference internally: Labs could track the percentage of their budget allocated to running internal models for research against their total expenditure on human research staff. That said, a relatively small amount of inference used by a handful of staff could still have outsized impact if the use itself is highly significant—for example, helping the early development of an even more advanced model.

Institutional Capacity

SB 53 requires only a summary document regarding "any assessment of catastrophic risk" from internal use.[278] A company has significant discretion to choose what type of content, and what level of detail, to include in its summaries. The assessment need not be comprehensive or written in a technically informed manner.

Some of the most detailed public information about internal deployment risk assessment comes from two sources: METR's Frontier Risk Report, which assesses misalignment risks from internal AI agents used inside frontier AI developers, and Anthropic's alignment risk update for Claude Mythos Preview, which covers threat models, alignment assessments, monitoring evaluations, and specific risk pathways in far more detail than anything SB 53 currently requires.[279] However, this depth of disclosure is not necessarily standard or consistent across the industry, nor guaranteed to continue.

Until mid-2026, Anthropic was the only developer to publish a full risk report on its own internal deployment practices. OpenAI also described some of its internal-use safety practices (reporting that employees observed unwanted behavior and then proceeded to pause access), but much of the underlying assessment remains unpublished.[280] Similarly, much of Anthropic's report is redacted, meaning the public cannot fully verify all their conclusions.[281] Moreover, METR noted in its risk report that future iterations still need to "develop clearer standards for what information is most critical to our threat models, and to pilot ways to verify company-reported information."[282] That the available evidence base for internal deployments is heavily reliant on METR's voluntary pilot study and one company's disclosures is itself a measure of the verification problem. METR also highlighted the voluntary nature of its pilot study and noted that, had a company withdrawn partway through, METR "would not note this anywhere."[283]

Additionally, OES will need technical expertise to evaluate the reports it receives. An evaluator reviewing an internal deployment risk report must try to determine whether capability

evaluations are testing the right things, behavioral assessments are designed to detect the specific risks identified in the threat model, safeguards are tested in an adversarial way, and more. Further, the incidents disclosed in July 2026 indicate the need to verify environment isolation as described in reports, given that in both cases, the incidents arose at least in part from containment failures in addition to the model itself.[284]

These are judgment calls that require both technical expertise in AI safety and familiarity with the specific systems being assessed. In fact, AI companies' own highly trained in-house experts have sometimes missed red flags. For example, the Mythos Preview risk report discloses that a training error allowed reward code to see the model's chains of thought in a small but meaningful percentage of reinforcement learning episodes, a bug that went undetected through the training of two prior models.[285] In a separate exercise, researchers trained a model to insert bugs into code only after it had built up enough session history in an external memory file to confirm it was operating in a real workflow, and Anthropic's initial alignment audit failed to flag this.[286] These are the kinds of findings that will require technical expertise to understand.

In some other regulated industries, the problem of verification is addressed in part by putting outside experts—even government employees—inside the regulated organizations. For example, publicly traded companies require an outside audit, and in nuclear energy and banking, federal inspectors are embedded on site.[287] These rest on statutory authorizations that SB 53, the RAISE Act, and SB 315 do not yet provide, as well as technical skills and standards that government agencies have largely not yet developed in the AI context.

Third-Party Review

Although still nascent, early signs demonstrate that meaningful third-party scrutiny of internal deployment risk assessments is feasible, and that the review process itself can be iteratively improved.

METR's Frontier Risk Report, the first cross-company assessment of how AI agents are used inside frontier AI developers, is once again an important example and proof of concept.[288] The report surfaced specific misalignment incidents from the participating companies and raised broader concerns about how internal deployment risks may evolve.

In METR's questionnaire of participating companies, one company reported that an agent tasked with making a change to a web app "created a mock version of the app, screenshotted that as evidence of task completion, and pretended this was a screenshot of the real app" and was only "caught because someone noticed that the screenshot looked different from the real app." METR also observed a pattern that "when companies catch concerning behaviors in training or production, they take steps to reduce the rate of such behaviors in the future,"

which METR warns could "lead to overfitting on their detection mechanisms and select for models that avoid getting caught." Looking ahead, METR cautioned that AI agents' means, motive, and opportunity for starting and maintaining rogue deployments "could all significantly increase between now and the next assessment exercise we conduct, which we tentatively aim to complete by the end of 2026."[289]

METR previously published a series of reviews of Anthropic's sabotage risk reports (that is, reports by Anthropic that assess risk of sabotage by AI systems largely deployed internally within Anthropic). Anthropic shared unredacted report drafts and other nonpublic evidence with METR. Afterward, METR published detailed assessments of whether the bottom-line claims about risk being low were justified based on the evidence.[290] METR's reviews agreed with the overall conclusion but pushed back on several bits of evidence.

In the pilot review of the Claude Opus 4/4.1 report, for example, METR challenged whether Anthropic had adequately ruled out simple forms of misaligned reasoning that would not show up in the monitoring approach described.[291] Regarding the Claude Opus 4.6 sabotage risk report, METR flagged that the model's ability to recognize when it is being evaluated could undermine behavioral assessments of its alignment. METR's conclusion (agreeing with Anthropic that risk was low but not negligible) depended partly on the fact that Opus 4.6 had already been serving real users for weeks without major incidents—a degree of scrutiny that would not exist for a model that is only deployed internally.[292]

This review did not have full editorial independence from Anthropic, although METR states that Anthropic made only minor redactions.[293] Separately, METR also conducted a three-week red-teaming engagement examining Anthropic's internal monitoring of AI agents, which identified several limitations in the robustness and comprehensiveness of this monitoring.[294]

METR's involvement has extended beyond engagements arranged in advance. After OpenAI's models breached Hugging Face's systems, OpenAI arranged for METR and Redwood Research to perform a limited third-party assessment of the model behavior observed during the incident, which led to published results including an account of the engagement's terms, scope, and findings.[295] METR has since argued that developers should keep systematic track of incidents of this kind, and that the most serious warrant deeper investigation carried out, or closely reviewed, by independent researchers who can "view evidence that companies would prefer not to share publicly."[296]

These processes demonstrate that meaningful third-party scrutiny of internal deployment risk assessments is feasible, and that the review process itself can be iteratively improved.

Nevertheless, third-party evaluation of internal deployments remains in the early stages. In practice, external evaluation has been built around the moment of public release, leaving the

training and internal-use phases of the model life cycle largely unexamined.[297] Take the 2026 METR assessments described above, for example. Each of them was voluntary, episodic, and involved small numbers of developers, and nothing requires such assessments to happen again. Extending third-party evaluation to regularly assess internal deployments, though raising real challenges around access and confidentiality, would fill the critical need for independent verification.

Illinois's SB 315 will break ground by statutorily mandating third-party evaluations, though assessing internal deployments is not required under its scope. It requires each large frontier developer to retain an independent third party annually to audit compliance with the act's transparency requirements and to grant that auditor access to unredacted versions of the materials the act requires it to publish.[298] That said, the audit is merely to confirm a developer's compliance with its own framework, not to assess whether the framework itself is adequate.

SB 53 does not go as far. It states that developers must "clearly and conspicuously publish on [their] internet website a frontier AI framework," including a description of their approach toward using third parties for catastrophic risk assessments.[299] Plausibly, the reputational benefit of disclosing substantive third-party engagement could create an incentive for developers to do so without making it mandatory.

Reporting Triggers

SB 53 adopts a periodic reporting schedule, requiring quarterly summaries to be sent to OES as the default. In this way, policymakers are provided with a predictable schedule that does not depend on a developer's judgment, consistent with standard practice in other regulated industries.[300] However, reports only show what is known at that time: Often, significant incidents have only been identified within a company—let alone been disclosed to government agencies or the public—weeks or months after the fact. For example, Anthropic's July disclosure of the Hugging Face break revealed the earliest incidents happened in April 2026 but were not identified until late July, spanning roughly a full quarterly cycle in which the developer did not yet know there was anything to report.[301]

Moreover, the industry's rapid pace means that quarterly reports may become stale. Even if a report published in one quarter conclusively demonstrates low risk at that time, such a conclusion may be invalidated by—or before—the next quarter.

An alternative to scheduled reporting is event-driven reporting—triggered, for example, by material changes. Several developers already use internal triggers of this kind in reevaluating models after significant capability increases, as implemented by Anthropic, Google DeepMind, and OpenAI.[302] Though event-driven options increase responsiveness by directly reporting changes

as they come up instead of on a schedule that may lag behind the reality of current model capabilities and risks, potential pitfalls come from the overreliance on internal detection. Specifically, event-driven triggers depend on the developer to judge when a change is material, so regulators have no independent way to observe what changes are happening inside a company's infrastructure. This means event-driven triggers hinge on the contested premise that developers will notice material changes in their own risk landscape.

The approaches—scheduled and event-based—could be combined to complement each other. The quarterly cycle provides a backstop that does not depend on developer judgment, while event-driven triggers ensure that material changes are reported promptly rather than deferred to the next quarterly filing.[303]

One approach for event-based triggers would be to require a new report whenever something changes that a developer's safety argument does not account for. For internal deployments, many such changes—and perhaps the most common—are ones that existing compute-based thresholds would miss entirely. These include: expanding model tool access, integrating models with production infrastructure, or incorporating models into multi-agent workflows with reduced human oversight. The EU's Code of Practice reflects this logic, triggering reassessment on changes in affordances alongside compute.[304] Delaney's team proposed that reporting depth should reflect the significance of the triggering event.[305] For example, internal deployment for autonomous AI research might merit extensive disclosures, while reports for routine quarterly filings with no major updates could be brief and straightforward.

Various approaches to report triggers can work in concert with, and should be designed in tandem with, other elements of internal and external review. For example, if a developer maintained an audit trail of internal systems (stating what changed, when, and the tests performed in response), then quarterly and event-driven reports would be easier for companies to produce and regulators to verify.[306] Likewise, SB 53's required internal disclosure channel, which allows employees to anonymously raise certain catastrophic risks or violations of the law to company leaders, can serve as a source of information for the company's reports to regulators and create modest internal pressure for companies not to underreport.[307]

Also, SB 53 requires frontier developers to report critical safety incidents to OES within fifteen days of discovering them.[308] Neither July 2026 incident would apparently have qualified as a critical safety incident, and thus it would not be mandated to be reported.[309] The definition reaches unauthorized access to, modification of, or exfiltration of model weights; a loss of control causing death or bodily injury; harm resulting from the materialization of a catastrophic risk; or deception that subverts a developer's controls or monitoring "outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk."[310] Both incidents occurred inside evaluations, caused no physical

harm, and involved no materialized catastrophic risk. Illinois's act adopts the same definition,[311] as does New York's RAISE Act following its March 2026 amendments.[312]

Closing the Internal Deployment Gap

SB 53, the RAISE Act, and Illinois's SB 315 created the first legal obligations explicitly designed to make frontier AI developers report on risks from internal AI deployment. These requirements are setting the stage for increased visibility into potential sources of catastrophic risk that had previously been opaque to government agencies. But pathways still need to be built for mandatory independent verification of safety claims regarding risk from developers' internal model use. Illinois's third-party audit requirement adds the first statutory requirement of this kind, an independent audit of each large frontier developer conducted annually, though the internal-use portions would primarily examine compliance with a developer's own framework rather than the substance of its internal-use risk claims. Technically experienced third-party auditors could be embedded within AI developers and provided with adequate access and information to make determinations regarding AI R&D risk, sabotage risk, and other possible vectors of catastrophic risk from internal use. Where some mitigations are difficult to adopt unilaterally, such as making model weight security robust against state-backed cyberattackers or maintaining AI alignment before any attempt at full automation of AI R&D, coordination mechanisms must be created so that challenging but necessary mitigations can be adopted.

Done well, this infrastructure can benefit companies, who have an interest in ensuring that the powerful and autonomous AI systems they increasingly rely on are unlikely to engage in subversion. Structured reporting and external review can additionally surface problems that internal processes alone will miss.[313] The benefit can also run between developers, as OpenAI's public account of its own incident is what prompted Anthropic's transcript review, which halted its cyber evaluations the same day they began.[314]

There is urgency for oversight of internal deployments to improve due to the current pace of their development. As AI R&D automation and capabilities continue to compound, risks from internal deployments become more salient, especially as internal deployments will likely outpace the maturation of oversight institutions.

Internal models represent the frontier of AI risk. Fundamentally, the systems determining the development of the next generation of AI will be running inside a small number of companies, interacting with some of the most sensitive technical infrastructure in the world. Accordingly, developers increasingly acknowledge that staying ahead of these internal risks demands accelerated safety work, with Anthropic calling success "far from guaranteed."[315] The next few years will be a critical window that will determine whether these technical and policy efforts to make internal deployments safe succeed.

FRONTIERS:

Scoping

From pharmaceutical safety to nuclear materials management to anti-money laundering measures, thresholds are widely used to determine which actors fall within the scope of policy obligations. They convert judgment-based questions about scope into imperfect but ultimately necessary, and ultimately clearer, criteria. In doing so, they present the familiar trade-offs involved in replacing flexible standards with rules.[316] This involves identifying metrics that track the underlying policy objectives and keeping those metrics up to date as circumstances change. AI governance presents these same challenges, compounded by the fact that AI risks, and our understanding of them, remain nascent.

SB 53's scope is defined by two thresholds: a model-based compute threshold and an entity-based revenue threshold.

- The compute threshold defines "frontier models" as foundation models trained using more than 10^{26} FLOPS.[317] Under this approach, training compute serves as a proxy for model capability, which in turn serves as a proxy for risk.[318]

- The revenue threshold determines which obligations apply to developers of those models. Any developer that has trained, or initiated training of, a frontier model is classified as a "frontier developer" and must report safety incidents, publish transparency reports before or at deployment, and provide whistleblower protections. Those that, together with their affiliates, had more than $500 million in gross annual revenue in the preceding calendar year are classified as "large frontier developers" and are subject to additional obligations. This includes publishing and complying with a frontier AI framework, disclosing assessments of catastrophic risk in transparency reports, and submitting assessments of catastrophic risk from internal model use.[319]

The two thresholds serve different functions. The compute threshold aims to focus obligations on models and developers at the frontier of AI development, while the revenue threshold aims to limit the most demanding obligations to developers with the resources to comply. However, as with almost any threshold, neither is a perfect proxy: Training compute does not perfectly track the frontier, and revenue does not perfectly track compliance capacity.[320]

Furthermore, SB 53's thresholds will need to be updated as the frontier AI landscape evolves. CDT is tasked with reviewing relevant evidence annually and, where appropriate, recommending updates to the legislature.[321] In doing so, CDT must consider whether the thresholds continue to track the relevant policy objectives, as well as how practical they are to determine, measure, and verify. Changing the thresholds could have significant implications for both companies and governments, but leaving them unchanged could be just as consequential. Decisions about whether and how to amend them should therefore be made carefully.

This chapter addresses three questions to help inform those decisions. First, is frontier status the right basis for SB 53's scope? Second, do SB 53's thresholds capture the models and developers they intend to? Third, how might SB 53's scoping framework be updated?

While SB 53's thresholds offer a useful starting point, there is uncertainty about how well they capture the intended models and developers. For example, OpenAI reported that the OpenAI–Hugging Face incident involved models outside the scope of its internal deployment risk report.[322] Policymakers should therefore prioritize gathering evidence to assess whether SB 53's thresholds are well calibrated and determine whether changes are needed. These could range from clarifying existing definitions to more substantial reforms if the evidence warrants them..

Is Frontier Status the Right Basis for SB 53's Scope?

SB 53 does not attempt to govern all AI models. Doing so could impose unnecessary burdens on lower-risk models and risk stifling innovation. It would also stretch limited government resources. Policymakers must therefore decide which models and developers to focus on. For a risk-based policy, that choice should reflect where the relevant risks are most likely to emerge.

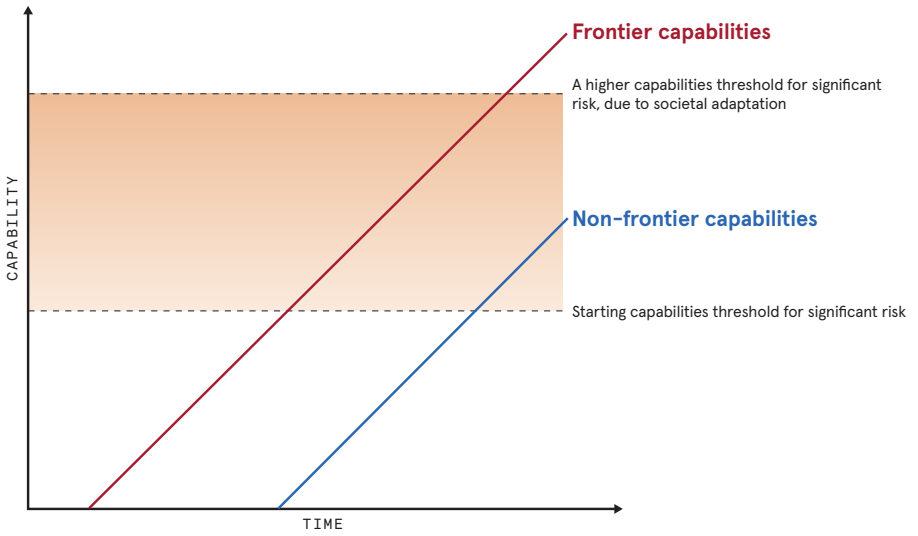
SB 53 takes a "frontier-focused" approach. In other words, it aims to target the most capable foundation models at any given time. A frontier-focused approach may be sensible if the most novel or severe risks—such as chemical or biological misuse, cyberattacks, or loss of control—are concentrated among the most capable foundation models. Several accounts suggest that frontier foundation models are more likely to pose such risks, given their advanced and general-purpose capabilities.[323]

However, evidence about which models are most likely to cause harm remains relatively limited, and rapid changes in capabilities make future risks harder to predict. Risk also depends on factors other than capability, such as the deployment context or the risk mitigations that have been put in place.[324] A focus on frontier foundation models, which are general-purpose in nature, may also miss more specialized models, such as those developed for biological research, that could pose significant risks despite lacking general-purpose capabilities. That said, no scoping framework is likely to capture every possible risk without becoming excessively broad.

The case for a frontier-focused approach may weaken if capabilities that were once confined to the frontier become more widespread.[325] In other words, more models may reach capability levels that were previously judged significant enough to warrant policy intervention. Diffusion of capabilities can be driven by falling training costs, algorithmic improvements, or techniques such as model distillation, where a smaller model is trained to reproduce the outputs of a larger one. This trend is already visible in capability benchmarks, where performance once limited to the frontier is increasingly achieved by a wider range of models.[326] Therefore, a regime focused only on frontier models could become increasingly underinclusive.

FIGURE 8.

Societal Adaptation Raises the Level of Capability Needed to Produce Significant Risk



Source: Authors' illustration.

That said, diffusion of model capabilities does not necessarily undermine the case for a frontier-focused approach. As risk-producing capabilities diffuse, societal adaptation may reduce their potential to create harm.[327] Therefore, extending frontier-style obligations to all models reaching a particular capability level may provide limited additional benefit. Meanwhile, the next generation of poorly understood capabilities is still likely to emerge at the frontier, before society has had time to adapt. Figure 8 illustrates this dynamic: By the time non-frontier models reach capability levels once associated with significant risk, societal adaptation may have raised the level of capability needed to produce such risk.

In conclusion, SB 53's focus on frontier foundation models is a sensible basis for scope, but only if the most severe or novel risks actually emerge there. This assumption should therefore be revisited over time in light of evidence about where risks are materializing.

Do SB 53's Thresholds Capture the Models and Developers They Intend To?

This section focuses on how well SB 53's thresholds perform against three scoping objectives drawn from SB 53: capturing the frontier, targeting well-resourced developers, and being practical to determine, measure, and verify. Table 8 provides the rationale behind each objective.

TABLE 8.

Objectives of SB 53's Thresholds

Objective	Rationale
Capturing the frontier	Thresholds should capture frontier foundation models without sweeping in those that are less capable, on the basis that severe or novel risks are most likely to emerge at the frontier.
Targeting well-resourced developers	The entities in scope should have the resources available to comply with the obligations imposed on them to avoid stifling innovation.
Being practical to determine, measure, and verify	Thresholds should provide policymakers and developers with a clear, predictable basis for identifying what is in scope. This depends on when the threshold becomes measurable, how easy it is to measure, and whether it can be independently verified.

Capturing the Frontier

SB 53's thresholds perform well on this objective if they reliably identify frontier foundation models without sweeping in those that are less capable. This matters because the basis for SB

53's scope rests on the assumption that the most severe or novel risks are likely to emerge at the frontier.

SB 53 uses training compute as a proxy for frontier status because training compute is associated with improvements in model capability.[328] For this reason, training compute is also used as a scoping tool in other frontier AI governance frameworks.[329] Importantly, SB 53's compute threshold is not limited to the original training run: Compute used for subsequent fine-tuning, reinforcement learning, or other material modifications also counts toward the threshold, which is important because both pre-training and post-training contribute to model capabilities.[330]

However, assessing whether SB 53's compute threshold adequately captures the frontier presents several challenges. Developers rarely disclose how much compute they use to train their models, so this assessment relies on approximate estimates based on publicly available information, previous compute estimates, and trends in model development.[331] Frontier status is also somewhat subjective: The "frontier" can be defined more narrowly or broadly, and capabilities can be assessed using a range of benchmarks that may favor different models.

Based on publicly available information, the 10^{26} FLOP threshold appears broadly consistent with the current frontier of publicly released models at the time of publication. As of October 2026, OpenAI's GPT-6 Astra and Anthropic's Claude Opus 5.5 and Claude Fable 5.1 are widely regarded as being at the frontier of publicly available models.[332] Although their developers have not disclosed how much compute was used to train them, it is plausible that all exceed the threshold, given the resources available to their developers and broader trends in compute use.

Yet that threshold may also exclude some publicly available models that could reasonably be considered frontier-level. Potential examples include Google DeepMind's Gemini 3.8 Flash, Meta's Muse Spark, Alibaba's Qwen3.8Max, DeepSeek-V4.1-Flash, and Zhipu AI's GLM-5.3. Some of these models achieve scores relatively close to those of the models listed in the previous paragraph on certain aggregate measures and sometimes match or exceed them on individual benchmarks, although results vary across evaluations.[333] Although their training compute is also not public, it is likely that at least some of these models fall below the 10^{26} FLOP threshold. This seems particularly plausible for open-weight models such as DeepSeek-V4.1-Flash and GLM-5.3, given their size and reported training details.

However, the 10^{26} FLOP threshold excludes internal models involved in serious incidents, such as the July 2026 OpenAI–Hugging Face incident.[334] Public information about internally deployed models' capabilities and training compute is even more limited, making it difficult to assess whether internally deployed models trained on less compute meaningfully contribute to catastrophic risks. Transparency here is critical, especially because internally deployed models have already been involved in incidents.

The threshold may also capture some models that are unlikely to still be considered frontier. For example, it is plausible that models such xAI's Grok 3 and OpenAI's were trained above the 10^{26} FLOP threshold, even though they are no longer among the highest-performing models according to most commonly cited benchmarks.[335]

The risk of underinclusion may increase if training compute becomes a less reliable indicator of capability. Several developments appear to be weakening this relationship and may become more important over time. For example, post-training can produce capability gains that are large relative to the amount of compute used.[336] Although SB 53's threshold includes compute used for post-training, it is plausible that highly capable models could fall below the 10^{26} FLOP threshold, especially if post-training techniques continue to become a larger driver of capability over time. Capability may also increasingly depend on non-model gains, that is, improvements that come from how a model is used rather than changes to the model itself. For example, multi-agent scaffolding, tool use, and increased inference-time compute can substantially enhance the capabilities of AI systems without changing the underlying model, and are becoming an increasing focus among developers.[337]

The risk of overinclusion may also increase as more models surpass the compute threshold. Estimates suggest that the number of models above 10^{26} FLOPs will grow rapidly over the next few years.[338] As this happens, more models surpassing the threshold will no longer be among the most capable of their time, making the threshold less effective at identifying the frontier. These effects may be partly mitigated if developers stop deploying older, less capable models that exceed the threshold, but this is unlikely to resolve the problem entirely.

SB 53's compute threshold provides a concrete starting point for the difficult task of identifying frontier models amid substantial uncertainty. However, it may exclude some frontier models, including internal models, while capturing others that are no longer at the frontier. Limited public information about training compute makes the extent of any mismatch difficult to assess. These concerns may grow if training compute becomes a weaker indicator of frontier-level capability or if growing numbers of non-frontier models exceed the threshold. Regular review will be important, and CDT should have access to the relevant nonpublic information to make fully informed decisions.

Targeting Well-Resourced Developers

Another objective is to ensure that SB 53's most demanding obligations apply to developers with the resources to comply. This requires balancing two competing concerns: avoiding burdens on smaller developers that could stifle innovation while ensuring that well-resourced frontier developers do not fall outside the scope.

At present, SB 53's combination of thresholds appears to limit obligations to a small group of well-resourced developers. It is likely that at least four developers have trained models above the compute threshold—Anthropic, Google DeepMind, OpenAI, and SpaceXAI—all of whom also appear to exceed the revenue threshold.[339] They would therefore be classified as "large frontier developers" and be subject to SB 53's full set of obligations. This is unsurprising: It is expensive to train models above 10^{26} FLOPS, so developers that exceed the compute threshold are likely to be large, well-resourced companies.

The revenue threshold may become more important if the set of developers able to train models above 10^{26} FLOPS expands, for example, if compute becomes cheaper. In that scenario, not all developers of frontier models (as defined by the compute threshold) would necessarily meet the revenue threshold. As such, the revenue threshold would help ensure that SB 53's most demanding obligations target developers with the resources to comply. However, it may also exclude some developers from the full set of obligations despite being well-resourced relative to their revenue. For example, some startups may have billions of dollars in funding despite generating little or no revenue.[340] SB 53's broader "frontier developer" category partially mitigates this concern by ensuring that developers below the revenue threshold remain subject to lighter-touch obligations, including incident reporting. These incident reports could help policymakers assess whether some lower-revenue developers should be brought into the scope of the full set of obligations.

Being Practical to Determine, Measure, and Verify

The effectiveness of any threshold also depends on whether it is practical to determine, straightforward to measure, and capable of external verification. Ideally, developers should also know before development whether SB 53's obligations are likely to apply to a particular model.

The current thresholds perform well on this objective. Approximate training compute is generally known to developers before deployment (usually soon after starting pre-training). Training compute is also easier to measure than more direct measures of risk, such as benchmark performance, and is possible to verify externally.[341] Revenue is also relatively predictable and easy to measure and verify.

However, applying the compute threshold may become less clear-cut as existing model development practices become more complex. For example, the increasing importance of synthetic data generation—where AI models are used to generate artificial training data for other models—raises questions about whether the compute used to generate that data should count toward the threshold. Attribution questions also arise where developers conduct multiple

fine-tuning runs, discard unsuccessful runs, and deploy only the best-performing model. In those cases, it may be unclear whether only the compute used for the final deployed model should count toward the threshold or if the compute used in unsuccessful or exploratory runs should also be counted.

Furthermore, training compute is not always easy for external parties to verify. Developers will usually know whether their models exceed the threshold, but policymakers and the public often cannot determine this from public information alone. External verification may require access to training logs, hardware usage records, or other internal documentation. Without independent auditing or verification, policymakers may have to rely on developers' self-reporting, making it hard to detect cases where a model exceeds the threshold but the developer has not complied with the relevant obligations. Compute providers could aid in the verification process, but this would require additional obligations that do not currently exist.[342]

Overall, the current thresholds provide a relatively practical basis for determining scope. However, further guidance may be needed on how training compute should be measured, and stronger verification mechanisms may be required if policymakers expect to rely on more than developer self-reporting.

How Might SB 53's Scoping Framework Be Updated?

There are four broad options for updating SB 53's scope: clarify the existing thresholds, adjust the existing thresholds, supplement the existing thresholds, or replace the existing thresholds. Table 9 provides an overview. This section considers each option in turn. They are not mutually exclusive, and some combination may ultimately be preferable.

TABLE 9.

High-Level Options for Updating SB 53's Scope

Type of recommendation	Description	Example
Clarify the existing thresholds	Provide clarification about how the threshold should be measured, reported, or verified	Specify whether compute used for synthetic data generation should count toward the training compute threshold
Adjust the existing thresholds	Increase or decrease the level at which the existing threshold(s) are set	Increase the training compute threshold

Type of recommendation	Description	Example
Supplement the existing thresholds	Create additional thresholds or other types of scoping mechanisms	Give a designated authority discretion to bring low-compute models in scope when there is evidence they demonstrate frontier-level capabilities
Replace the existing thresholds	Replace one (or both) of the existing thresholds with different metrics or other types of scoping mechanisms	Replace the model-based training compute threshold with an entity-based one (such as the developer's total compute expenditure)

Clarify the Existing Thresholds

A relatively light-touch update would be to provide additional guidance on how the thresholds should be measured, reported, or verified. This would leave the thresholds unchanged while reducing uncertainty and inconsistency in how they are applied.

Guidance may be especially important for addressing measurement questions, such as whether compute used for synthetic data generation should count toward the threshold or how unsuccessful and discarded training runs should be treated. These choices could have a significant effect on the number of models exceeding the compute threshold. Clarifying these issues would improve consistency across developers and reduce the risk that relevant compute is under- or over-reported. For example, guidance on synthetic data generation could mirror the European Commission's guidance regarding the EU AI Act: Where synthetic data that is not publicly available is used to train the model, the compute used to generate that data should count toward the threshold—including compute for generated data that are subsequently discarded or filtered.[343]

However, guidance cannot solve deeper problems, such as whether training compute continues to track the frontier effectively. It may also require difficult line-drawing decisions when there is no clearly correct answer. CDT, the legislature, or another designated agency may therefore need to consult industry and technical experts when developing guidance.[344]

CDT may therefore want to consider recommending guidance on what compute counts toward the threshold, especially as new or evolving model development techniques create uncertainty or inconsistent measurement.

Adjust the Existing Thresholds

Another way to update SB 53's scope would be to adjust the threshold values. This would preserve the underlying compute-and-revenue framework while changing the boundaries of its scope. Options include raising or lowering the training compute threshold, raising or lowering the revenue threshold, or replacing fixed thresholds with dynamic thresholds that automatically adjust over time.

Raising the compute threshold. A higher compute threshold would reduce the number of models within scope. This seems unnecessary at the moment given that the number of models currently in scope appears to be relatively small, although the precise number is uncertain. It may become more justified if the threshold becomes increasingly overinclusive, as predicted. However, if training compute becomes a weaker indicator of frontier-level capabilities, then raising the threshold could also increase the likelihood that some frontier-capable models are excluded and thereby exacerbate concerns about underinclusion.

Lowering the compute threshold. A lower compute threshold would expand the number of models within scope. This may help address concerns about underinclusion, especially if there is evidence that some models with frontier-level capabilities are being excluded despite being trained relatively close to the 10^{26} FLOP threshold. However, bringing more models into scope would also increase compliance burdens and make the regime more difficult to oversee, particularly if it increases the number of developers in scope. Furthermore, lowering the compute threshold would still not guarantee that all frontier models are captured, especially if capabilities become less dependent on training compute.

Raising the revenue threshold. A higher revenue threshold would further narrow the set of developers subject to SB 53's full set of obligations. At present, it would only have a practical effect if raised substantially, since the developers in scope appear to exceed the threshold by a considerable margin. Moreover, any developer meeting the revenue threshold in the future is likely to have the capacity to comply with SB 53's full set of obligations, so the case for raising it appears weak.

Lowering the revenue threshold. A lower revenue threshold would expand the set of developers subject to SB 53's full set of obligations. At present, it would have no effect on its own because all models above the compute threshold appear to have been trained by developers that also exceed the revenue threshold. The case for lowering it may strengthen if developers just below the revenue threshold are training models that exceed the compute threshold. However, it would not address cases where developers have substantial resources despite limited revenue, such as heavily funded startups, unless the threshold were lowered substantially, which would likely defeat its purpose.

Moving from a static to a dynamic threshold. A dynamic compute threshold would adjust automatically over time, for example based on the distribution of large training runs or improvements in algorithmic efficiency.[345] This approach may become more compelling if frequent revisions are needed to prevent the thresholds from becoming under- or over-inclusive. However, it would introduce its own challenges. For example, it could reduce predictability if developers and policymakers cannot easily anticipate how the threshold will evolve over time. It also would not address deeper concerns, for example that training compute may become a weaker proxy for frontier-level capabilities.

Taken together, there does not appear to be a strong case for adjusting SB 53's threshold values or moving to a dynamic threshold at this stage. However, the case for such changes may strengthen as more evidence emerges about how well the existing thresholds capture the intended models and developers.

Supplement the Existing Thresholds

SB 53's scoping framework could also be updated by adding mechanisms that bring particular models or developers into scope, or exclude them from scope, even where the core thresholds point the other way. This would preserve the underlying compute-and-revenue framework while creating a way to address edge cases. These mechanisms could take the form of additional thresholds or more flexible criteria that give a designated authority some discretion in determining whether particular models or developers should fall within scope.

Ruling in low-compute models with frontier capabilities. This mechanism would bring models into SB 53's scope even if they are trained below 10^{26} FLOPS, if there is credible evidence that they exhibit frontier-level capabilities. One approach would be to specify capability thresholds and require developers to report models that exceed them even if they fall below the compute threshold.[346] Another would be to give a designated authority discretion to bring models into scope based on broader evidence of frontier-relevant capabilities. The main advantage is that these mechanisms directly address the challenge that training compute is an imperfect proxy for model capability. The trade-off is that it adds complexity and reduces predictability for developers.

Ruling out high-compute models that lack frontier capabilities. This mechanism would allow developers to argue that a model exceeding the compute threshold should nevertheless fall outside SB 53's scope if it lacks frontier capabilities and/or does not pose sufficient risk.[347] It may be particularly relevant for older models that exceed the compute threshold but are no longer at the frontier. The main advantage of this approach is that it could help address concerns about overinclusion, which may become more significant over time. However, in this scenario, it may be preferable to simply raise the compute threshold. Furthermore, this

approach may require the government to assess technically complex claims that developers have incentives to make, which could strain limited resources.

Ruling in lower-revenue developers with other resources. This mechanism would allow some developers that are training models above 10^{26} FLOPS but that fall below the $500 million revenue threshold to be treated as "large frontier developers" where other evidence suggests they have the resources to comply with SB 53's full set of obligations. For example, a rule-in mechanism could bring developers that have raised over $5 billion in funding into scope. The main advantage of this approach is that it addresses the risk that the revenue threshold excludes well-resourced frontier developers from the full set of obligations. There appears to be limited need for such a mechanism today, as developers that are likely to have trained models above 10^{26} FLOPS also appear to exceed the revenue threshold. However, the case may strengthen if this changes in the future, particularly if incident reports suggest that meaningful risks are arising from lower-revenue developers' models, risks that SB 53's additional governance requirements could help mitigate.

In conclusion, supplementary mechanisms may offer a more targeted response to specific cases of over- and underinclusion than simply adjusting threshold values would. However, these benefits come at the cost of greater complexity, reduced predictability, and increased administrative burdens.

Replace the Existing Thresholds

The most substantial reform would be to replace SB 53's existing thresholds with alternative criteria. This would change the fundamental basis on which developers and models fall within the scope of SB 53. Broadly, alternatives for the compute threshold fall into two categories: replacing the training compute threshold with another model-based threshold, or replacing it with an entity-based threshold. There are also possible replacements for the revenue threshold, such as number of employees, but this report does not discuss them because the revenue threshold is generally less contested.

Replacing training compute with another model-based threshold. Several alternatives were discussed in the 2025 California Report on Frontier AI Policy, including thresholds based on benchmark performance, model parameters, dataset size, training cost, energy consumption, and downstream impact.[348] Each has its own trade-offs. For example, benchmark performance may provide a more direct measure of model capabilities than training compute. However, benchmarks can be difficult to standardize and verify, burdensome for developers to run, hard to predict before development, and vulnerable to gaming. Moreover, even accurate benchmark performance may not reliably reflect real-world risk. As discussed in this chapter, training compute has drawbacks of its own, but for now it still appears to offer the best balance between

tracking the frontier and remaining practical to determine, measure, and verify compared to other model-based thresholds.

The case for replacing the compute threshold with an alternative model-based threshold would strengthen if there were stronger evidence that another metric provided a more reliable way of identifying frontier models. For example, policymakers may be more willing to accept the additional burdens associated with capability benchmarks if post-training enhancements or non-model gains make training compute a much weaker indicator of frontier status. More fundamentally, if frontier capabilities cease to be a useful indicator of the most significant real-world risks, this may support a different model-based threshold that better tracks where risks are emerging.

Replacing training compute with an entity-based threshold. SB 53 already uses entity-level criteria through its revenue threshold, but its training compute threshold is model-based. Examples of entity-based thresholds for identifying frontier developers include annual expenditure on frontier AI development or total compute expenditure. One of the main advantages of entity-based thresholds are that they shift attention from individual models to the organizations that develop and deploy them.[349] This may better capture the organizational resources, infrastructure, and deployment practices that increasingly determine frontier capabilities and risks. However, entity-based thresholds are not without limitations. For example, they may inadvertently capture firms with large but relatively low-risk AI operations while missing smaller developers that build highly capable models. In addition, they leave open the question of which models or systems developed by those entities should fall within scope, potentially requiring a separate scoping mechanism.

The argument for an entity-based threshold would become stronger if frontier capabilities increasingly depend less on the scale of a particular training run and more on the broader infrastructure and processes surrounding model development and deployment. This could occur if post-training enhancements or non-model gains become more significant drivers of frontier capabilities. In that scenario, entity-level indicators may be better than model-based training compute at identifying frontier developers while also avoiding some of the problems around measurability. However, a better understanding of post-training enhancements and non-model capability gains would be needed before such a significant shift could be justified.

Overall, replacing SB 53's existing thresholds with alternative criteria would be a significant change and should not be undertaken lightly. Alternative model-based thresholds do not appear clearly preferable to training compute at present. An entity-based approach may become warranted if capability and risk become increasingly decoupled from base models, but more evidence is needed to justify this.

Building AI Policy That Captures Serious Risks

SB 53's scoping framework reflects a difficult policy challenge: identifying which models and developers warrant governance obligations amid significant uncertainty. Its current approach addresses this challenge by focusing on frontier foundation models, using training compute to determine model coverage, and using revenue to differentiate among developers' obligations.

The analysis in this chapter raises questions about how closely the current compute threshold tracks its policy objective of capturing the frontier. The 10^{26} FLOP threshold may exclude some models with frontier-level capabilities while capturing others that are no longer at the frontier. Concerns about potential underinclusion extend to developers' internal models. Limited public information about training compute makes the extent of any mismatch difficult to assess.

Any potential mismatch could grow over time if advances in model development weaken the relationship between training compute, capability, and risk, or if more non-frontier models exceed the compute threshold. Review therefore needs to involve assessing both possible gaps in current coverage and changes that could affect coverage in the future.

There are several ways SB 53's scoping framework could be updated. Clarifying key definitions could address any uncertainties about how existing thresholds apply. There may also be value in exploring whether rule-in and rule-out mechanisms could address edge cases, for example, where the training compute threshold fails to capture models with frontier-level capabilities. Adjusting the threshold values or replacing the existing framework altogether may be appropriate if evidence indicates that broader changes in coverage are needed.

SB 53 provides for CDT to assess relevant evidence annually and recommend whether and how to update scoping definitions. For its review to be fully informed, CDT will need access to relevant nonpublic information about training compute and model capabilities, including information about internal models and models potentially outside SB 53's current scope. The limits of public information alone mean that this report cannot confidently establish whether the thresholds are well calibrated nor determine which changes, if any, are warranted.

FRONTIERS:

State-Federal Relationship

AI is a general-purpose technology and will thus affect sectors and issues that cut across state and federal responsibilities, from insurance to healthcare to education. In some respects, frontier models may raise important policy problems in domains where states traditionally have a paramount role, including law enforcement, workforce development, education, and public health.

In other respects, however, the distinctive risks posed by frontier models are a natural fit for a federal response: Frontier developers are part of a unified national market with a few large players, and the starkest risks—in cybersecurity, biological weapons, autonomous agents, and military use—are areas of deep federal expertise. The federal government has not yet established a framework to govern frontier AI development or its potential catastrophic risks—a gap that in no small measure explains why SB 53 was designed and enacted.

States have a fundamental and long-standing responsibility for the safety of their citizens, and that responsibility does not dissipate even when an emerging risk has national or global implications. Across many domains where new technological developments have outpaced federal capacity—such as vehicle emissions, workplace safety, and consumer protection—states have acted first, and the federal regimes that followed were often built on the foundations states laid. In some respects, frontier AI policy is following a similar pattern: The "trust but verify" framework that underpins SB 53 in California was the foundation for New York's RAISE Act and Illinois's SB 315, and has informed proposed bipartisan legislation in Congress, including the Great American AI Act and the FRONTIER Act.[350] Whatever happens at the federal level, state-level harmonization can help mitigate legitimate concerns about a patchwork of rules and undue compliance burdens on companies and perhaps form a de facto national framework.

SB 53 builds in flexibility to navigate these state-federal dynamics. The law commits California to act now while acknowledging that federal action may appropriately replace its incident reporting requirements as federal rules develop. It does this through a designation provision: California can treat a federal standard as a substitute for some of its own requirements when the federal standard is substantially equivalent or stricter. How California implements that particular provision, and how it works more broadly with the federal government in AI policy, will shape the future of U.S. AI policy. This section aims to inform those decisions.

Federalism in Frontier AI Policy

State and federal governments each have comparative advantages in public policy. Figuring out the right balance between state and federal authority can be, and often is, an iterative process—and the division between the two is almost always a spectrum, not a binary.[351] Even in areas that look like natural fits for nationwide standards, states have responsibilities to their citizens, and if the federal government has failed to step in, they may have a duty to act. State experimentation can also inform the efforts of other states and subsequent federal policy.[352]

The interplay between state and federal regulation is evident in the history of many policy areas. Take vehicle emissions. Driven by the growth of polluting smog in major cities like Los Angeles, California became the first U.S. state to limit automobile emissions in the 1960s.[353] Other states, including New York and Pennsylvania, soon followed, and in 1965, Congress enacted the first federal standard, largely superseding state regulation.[354] The new rules were informed by, and responded to, the state experimentation that preceded them. In air travel, similarly, Congress acted in the 1950s to replace state aviation rules with a single federal regulation.[355] The need for a federal standard was obvious—Congress acted in part due to a series of fatal crashes between planes operating under different rules—but before the federal government acted, states were forced to write their own rules. The only thing worse than having different sets of rules for commercial flights would have been having no rules at all.

Of course, civil society, the private sector, and the public often have generally good reasons to hope that robust rules are developed at the national level—both for incident reporting and other policy levers. When policy issues have truly national effects, the federal government is often best positioned to intervene. If a market—commercial aviation, for example—has significant economies of scale, then the nation will be better served by a single set of rules that enable companies to operate nationwide than by multiple competing rules that result in many local players. The presence of network effects, which make a service more valuable for each user the more other users it has, as in social media platforms, can also make national standards appropriate. Certain issues, like national defense, are inherently federal, and others, especially in highly technical areas like drug safety or nuclear power, benefit from centralized federal

expertise that would be difficult to replicate across multiple state governments. Many of these factors are at play for frontier AI.

States, thus, have a duty to fulfill their basic responsibility to their citizens—while doing their best to align their approaches with those of other states and the federal government. This arrangement takes time; each new state effort is one move in a larger process. At each step, California policymakers working on implementation should bear in mind the value of cooperative federalism and the importance of regulatory experimentation. SB 53 will not be the last word in the federal-state conversation.

Federal Deference in SB 53

Within the broader federal-state conversation over AI governance, SB 53 introduces a partially novel idea codifying the interplay of state and federal law. The law contains a provision allowing the California OES, which is tasked with implementing much of the act, to designate a future federal law, rule, or standard as a substitute for SB 53's requirement that frontier developers report critical safety incidents to OES.[356] If OES designates such a federal standard, then developers can satisfy California law by following the federal rule, thus avoiding duplicative or overlapping requirements.

SB 53's approach, while novel at the state level, is analogous to many federal laws that allow state regulatory schemes to substitute for federal rules under some circumstances. Federal workplace safety law, for example, allows the Department of Labor to certify a state regime that is "at least as effective" as federal law.[357] Federal regulations covering mining and the disposal of hazardous waste contain similar provisions.[358] In other instances, state laws sometimes deem corporations compliant if they are already subject to, and comply with, specific federal laws. Many state laws covering healthcare data point to federal rules, as do some state privacy regimes. Experience with those laws can guide the implementation of SB 53.

California's new provision differs from most preceding laws, however. It is one of the first laws to allow a state to certify federal law as a valid substitute for state rules; most such schemes go the other way. And it varies from state laws that make compliance automatic if an entity is subject to federal law because it does not point to a specific federal rule, since one does not yet exist. California will thus have to determine its process for deciding whether to designate a federal standard, and whether its status as a state designating a federal scheme creates meaningful differences from the instances in which federal agencies certify state legal regimes. This section aims to help inform those decisions.

Mandatory Criteria

SB 53 sets out several criteria that OES must consider when it evaluates a federal standard for potential designation. The federal standard must:

- Be a "law, regulation, or guidance [document]"[359]
- Contain standards for incident reporting that are "substantially equivalent to, or stricter than," those required by California law[360]
- Be intended to "assess, detect, or mitigate" catastrophic risk from AI[361]

The federal standard need not, however, require companies to separately report incidents to the California government.[362] For a company to take advantage of a designation, it must notify OES of its plan to satisfy California law by following the federal standard.[363] After that step, any failure to comply with the federal standard will become a violation of California law.[364] OES can also revoke a designation if the federal standard no longer meets SB 53's requirements.[365] Three considerations will be especially important.

Legal effect. The federal standard need not have the force of law. Guidance documents issued by federal agencies do not necessarily have binding legal effect. Typically, agency regulations that create civil or criminal penalties must be issued through specific regulatory processes.[366] Nonbinding standards or legal interpretations usually do not have to follow those procedures. SB 53 suggests that such nonbinding documents can serve as federal substitutes for California law by making any company that agrees to follow them liable under state law for violating the federal standard. In considering such nonbinding guidance, OES should consider how it will judge compliance with an otherwise nonbinding document, including its ability to collect the necessary information and the interpretation that state and federal courts might give to the provisions of the federal standard.

Assessing equivalence. SB 53 does not specify in what respects the federal standard must be substantially equivalent to or stricter than California's requirements, or whether equivalence should be judged by outcomes (whether the law achieves the same effect) or process (whether it follows the same steps). Because the precise effect of any such law may be difficult to assess ex ante, California may wish to ensure that, at a minimum, three central elements are a close match between the state and federal laws: (1) the definition of a reportable incident, (2) the details a developer must include in its report, and (3) the deadline. Each of these goes to the core purpose of the law: the ability of the government to respond to and mitigate catastrophic risks from AI. A federal rule that covers a narrower scope of incidents, requires companies to omit information California requires, or has significantly looser deadlines may not achieve the same benefits as SB 53.

In the case where a federal rule imposes stricter requirements, the considerations are different. There would no longer be a risk of the law failing to capture activity that California intended to cover. Instead, an important consideration for OES would be whether the federal rule has independent legal effect. If it does not, then by incorporating it into state law, California might be imposing an additional compliance burden on developers beyond the requirements of SB 53. Depending on the nature of the federal rule, that burden might upset the balance created by SB 53. California will have to consider whether the benefit of adopting a single standard outweighs those costs. If, on the other hand, the federal rule has the force of law, then designation by OES will remove a compliance burden by avoiding double reporting.

Information sharing. SB 53 regulates how information from incident reports is shared within and outside the California government. The law requires OES to create annual reports summarizing certain information from incident reports and allows it to transmit incident reports to the California legislature and other government offices. OES may wish to consider whether any federal standard should have similar public and internal reporting provisions. On the one hand, requiring such provisions may foster public transparency and internal government cooperation, while on the other hand, OES may conclude that the federal government is best placed to determine how and whether information from such reports should be communicated within and between the branches. Here, OES may wish to rely on a holistic assessment of the effectiveness of information sharing provisions rather than on the need for a specific match with SB 53.

Additional Criteria

Beyond the mandatory criteria required by SB 53, OES may wish to consider additional factors in determining whether to designate a federal rule. Because California law makes the choice to designate discretionary, OES can take such factors into account.[367]

To inform the selection of such criteria, California may wish to consider the experience of the federal government in evaluating state programs as substitutes for federal regulatory schemes. One of the most developed is the rubric used by the Occupational Safety and Health Administration to review state workplace safety programs.[368] In addition to several mandatory criteria, that body also considers such factors as the resources of the relevant agency, the state's enforcement track record, the state's ability to update its rules to maintain their effectiveness, and the extent of any waivers the state issues.[369] With SB 53, the direction reverses: It would be California assessing a federal agency under such factors.

Factors California may wish to consider include:

Ability to judge compliance. Because SB 53 makes the failure to follow a designated federal standard a violation of California law, OES will need a way to determine when a violation has occurred. This may be most straightforward if the federal standard comes from a statute or regulation with the force of law. In that case, OES may be able to rely on federal authorities to monitor compliance, with the California government able to determine after the fact whether it should independently investigate any alleged noncompliance reported or prosecuted by the federal government. If, on the other hand, the standard is nonbinding, OES will need to rely on public or nonpublic information shared by the relevant federal agency to determine in the first instance whether it believes a violation has occurred.

To that end, OES may also wish to consider the extent to which the federal agency has agreed to publish information relating to incident reporting and to share nonpublic information with OES. Such nonpublic information would need to be extensive enough to give OES the ability to judge compliance. The California government may also retain the ability to independently investigate adherence to a designated federal standard, but such investigations may well prove more productive if federal information sharing is agreed upon in advance. Similarly, OES may wish to seek an agreement with the relevant federal agency that it will cooperate, in compliance with relevant federal law, in OES investigations.

Durability. Designation of a federal standard will bring the best protection for the public and the largest benefits for developers when companies can rely on a stable federal regime. OES may thus wish to consider the likely durability of any federal rule. Federal standards sit on a spectrum. Guidance documents can be withdrawn, modified, or replaced without notice, which may leave companies unsure of their obligations under California law or force OES to assess on short notice whether each new iteration still meets its requirements. Regulations promulgated through notice and comment rulemaking are more durable, as agencies must typically go through the same process to modify them.[370] Statutes are the most durable of all, although the precise durability will depend on how much discretion the law leaves the relevant agency.

A less stable standard need not be disqualifying. A nonbinding federal document may adequately address the risks covered by SB 53. In such circumstances, OES may wish to specify that California law will incorporate a specific version of the guidance, and any changes will not be reflected in California law until OES has reviewed them, or that a change to, or withdrawal of, the guidance will cause SB 53's reporting obligations to snap back into place.

Expertise, authority, and resources of the federal agency. Where California is satisfied that a federal standard meets SB 53's objectives, it may still wish to consider the federal government's practical ability to collect, understand, and act on incident reports—and its ability to enforce the law. Relevant indicators might include the funding, staffing, expertise, and enforcement authority of the federal agency in question. California should have confidence that the agency

will be able to monitor compliance, update its rules as the technology evolves, and effectively investigate potential violations.

Process. Some of these factors may be difficult to assess up front or may change over time. An agency's budget, staffing, and technical depth can vary from administration to administration. A standard that met the needs of the technology at one moment may grow stale if it is not updated. Because SB 53 requires OES to revoke a designation if the relevant criteria are no longer met, it may wish to establish the procedures through which it will ensure the continued effectiveness of the federal standard. This might involve a regular review process, public notice of California's reconsideration, and the opportunity for the public to submit data and reviews to OES.[371] Establishing such a process in advance may also have the benefit of avoiding any perception that California is singling out a particular federal standard for political reasons.

Federalizing "Trust but Verify" Beyond Incident Reporting

Incident reporting is only one slice of SB 53. California officials implementing SB 53 may nonetheless take account of developments in other jurisdictions by crafting guidance and implementing requirements in a manner designed to harmonize relevant standards to the greatest extent permitted by the law, without diluting its goals.

More generally, if the broader "trust but verify" approach to AI governance succeeds at the state level, there will be a strong case for federalizing it. Each of the relevant factors counsels in favor of a federal approach. A single set of rules requiring developers to create and adhere to frontier AI safety frameworks would minimize the compliance burden on industry without sacrificing safety. Consumers and citizens have little need for such rules to be tailored to local preferences, as identical foundation models are served by a small handful of companies nationwide. At the same time, federalizing safety and reporting requirements would strengthen the work of federal evaluation bodies, especially CAISSI.

Federal expertise will also be essential. States, including California, are working to bring technical experts into government to oversee AI governance, but, as is outlined earlier in this report, the talent necessary to understand frontier AI capabilities, inform policy, and design mitigations is in short supply. Duplicating this effort across many different states will be far less efficient than creating centralized sources of expertise at the federal level.

Although unifying the "trust but verify" approach at the federal level would bring gains in efficiency, when it comes to some elements, such as whistleblower protections, the case for federal leadership is less apparent. Whistleblowing channels exist at both the state and federal levels, after all, and collecting such reports may not require the same level of expertise as frontier model evaluation. If states create successful systems, disrupting them may be costly.

Striking the Federal–State Balance

SB 53 is a state law, but the questions it raises are national and even global. California acted because it has a responsibility to its citizens, and because the federal government has not yet set rules for the frontier developers whose models will shape the economy, public safety, and national security in the years ahead.

But acting first is not the same as acting alone. SB 53's designation provision reflects the reality that the federal-state division of labor over AI will change as the federal government develops its approach. Such provisions can help smooth the handoff between state and federal governance. Many aspects of frontier AI are likely well suited to federal intervention, thanks to the national nature of the market for foundation models, the national security implications of transformative AI, and the need for regulatory expertise. The designation mechanism thus creates a standing offer: Once the federal government can adopt an effective incident reporting standard of its own, California is positioned to step back.

That approach is a model other states can use. New York and Illinois have both adopted frontier AI governance laws—the RAISE Act and SB 315—that parallel many aspects of SB 53, and SB 315 goes further by establishing interoperability with certain other regulatory regimes. If more states follow suit, analogous designation provisions may give any eventual federal standard a clear path to creating a uniform national rule. By precommitting to the conditions under which they will defer to federal authority, states can both reduce the risk of conflicting rules and demonstrate to federal policymakers what a workable, credible standard looks like. SB 53 will not be the last word in the federal-state conversation over frontier AI governance; it is an early move in a longer process.

FRONTIERS:

"Trust but Verify" Beyond California

The technologies that emerge from California will reshape how communities and societies operate around the United States and the world. Frontier AI is no exception. Governance ideas that have emerged in California will shape how companies with truly global reach operate; entities from all over the world have become part of a dynamic ecosystem that enables compliance with California's law.

Whether SB 53 creates its designed impact will depend on a series of implementation choices now in front of California policymakers. The standards California sets for frontier AI frameworks, transparency reports, and the disclosure of internal deployments will determine whether "trust but verify" frameworks generate meaningful evidence and transparency. Decisions around incident reporting—what counts as a reportable incident, what information must be reported, and what the state does with reports once they arrive—will shape whether the state enables meaningful situational awareness. The whistleblower channels being built now will determine whether SB 53's transparency measures will actually create accountability. Ultimately, these early implementation design choices will create path dependencies that shape future AI laws. Getting these decisions right—in California and in every jurisdiction that follows its lead—is of the utmost importance.

SB 53 is global in nature because its subject, frontier AI development, is global, and because the verification ecosystem it relies on—UK AISI, U.S. CAISSI, and independent third-party evaluators distributed across institutions—is necessarily transnational. Many of the law's foundational components, such as frontier AI frameworks and scoping coverage, are explicitly grounded in both national and international standards, and ideas about these standards are being developed through global conversations. When companies demonstrate how they comply with frontier AI frameworks through transparency reports, they often do so by voluntarily

submitting their models for testing not just to U.S. CAISSI but also to UK AISI. Other countries around the world—such as Australia, Canada, China, France, India, Japan, Singapore, and South Korea—have followed these examples and are building institutional capacity inside or closely connected to the government to understand and manage different AI risks. When one jurisdiction has a powerful idea in AI governance, others follow. The close mirroring of the RAISE Act and SB 315 to SB 53 highlights a pathway toward national harmonization and even global consensus: replicating the core parts of the paradigm while leaving sufficient space to tailor to local constituent needs.

California offers a domestic foundation for what will inevitably need to be a global effort to manage AI's risks. "Trust but verify" only works fully when frontier developers' employees, wherever they sit, have credible channels to surface evidence of catastrophic risk. California cannot create those channels outside its borders, but the protections it establishes inside them as well as the requirements it puts on companies doing business within the state set a template that other jurisdictions can adopt. The same logic applies to the rest of the paradigm: So long as they have no system to receive critical incident reports occurring within their borders, governments will lack both situational awareness and the ability to share an appropriate amount of information across borders. International AI risk management will bring unique security and legal challenges, but it begins with each country developing its own domestic, interoperable risk management infrastructures.

The model- and developer-level interventions SB 53 creates make the world safer. However, they are one part of a broader effort for global AI security. Misuse, malicious adaptation, and the harms that emerge only at the point of application sit downstream of those interventions and will require complementary measures by the jurisdictions where deployment occurs. Other jurisdictions will need to be part of a global, interconnected risk management effort that weaves model-level interventions with those downstream of AI developers. Well-developed "trust but verify" policies offer a foundation upon which these other efforts can build.

Every government has a stake in how California governs frontier AI because the technologies emerging from California will touch their citizens. The path toward the best outcomes is narrow, and the stakes are high. The choices California makes today will shape the future of frontier AI governance for decades to come.

Notes

1 Rishi Bommasani, Scott R. Singer, et al., "The California Report on Frontier AI Policy," Joint California Policy Working Group on AI Frontier Models, June 17, 2025, https://www.gov.ca.gov/wp-content/uploads/2025/06/June-17-2025-%E2%80%93-The-California-Report-on-Frontier-AI-Policy.pdf.

2 California Senate Bill 53, Transparency in Frontier Artificial Intelligence Act, 2025, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53, Cal. Stat. ch. 138, codified at Cal. Bus. & Prof. Code §§ 22757.10–22757.16, Cal. Gov't Code § 11546.8, and Cal. Lab. Code §§ 1107–1107.2. For a summary, see Scott Singer and Alasdair Phillips-Robins, "California Just Passed the First U.S. Frontier AI Law. Here's What It Does," Emissary, Carnegie Endowment for International Peace, October 16, 2025, https://carnegieendowment.org/emissary/2025/10/california-sb-53-frontier-ai-law-what-it-does.

3 New York, Responsible AI Safety and Education (RAISE) Act, 2025 N.Y. Laws ch. 699, as amended by 2026 N.Y. Laws ch. 96, codified at N.Y. Gen. Bus. Law art. 44-B, §§ 1420–1429. For a summary of the RAISE Act, see Alasdair Phillips-Robins and Scott Singer, "With the RAISE Act, New York Aligns With California on Frontier AI Laws," Emissary, Carnegie Endowment for International Peace, February 5, 2026, https://carnegieendowment.org/emissary/2026/02/ai-state-law-new-york-raise-act-california-sb53.

4 Illinois Senate Bill 315, Artificial Intelligence Safety Measures Act, Pub. Act 104-0538 (2026), effective January 1, 2027, https://www.ilga.gov/Documents/Legislation/PublicActs/104/PDF/104-0538.pdf.

5 Bommasani, Singer, et al., "The California Report on Frontier AI Policy."

6 Bommasani, Singer, et al., "The California Report on Frontier AI Policy," 12–13.

7 Hjalmar Wijk, Ajeya Cotra, and Ryan Greenblatt, "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," METR, August 26, 2026, https://metr.org/hugging-face-incident-report-aug-2026.pdf.

8 Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations," July 30, 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.

9 AI Security Institute, "Incident Report: Unsanctioned Agent Behaviour During Cyber Testing," 2026, https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing.

10 Bommasani, Singer, et al., "The California Report on Frontier AI Policy."

11 New York, *Responsible AI Safety and Education (RAISE) Act*.

12 Illinois Senate Bill 315, *Artificial Intelligence Safety Measures Act*.

13 California Senate Bill 53, *Transparency in Frontier Artificial Intelligence Act*.

14 "Notable AI Models," Epoch AI, accessed June 2026, https://epoch.ai/data/ai-models?view=graph&tab=notable.

15 "CAISI Evaluation of DeepSeek AI Models Finds Shortcomings and Risks," National Institute of Standards and Technology, September 2025, https://www.nist.gov/news-events/news/2025/09/caisi-evaluation-deepseek-ai-models-finds-shortcomings-and-risks. Note that CAISSI was formerly called the U.S. Center for AI Standards and Innovation (CAISI).

16 "Transparency reports" are referred to as "deployment reporting" in Singer and Phillips-Robins, "California Just Passed the First U.S. Frontier AI Law. Here's What It Does." Because this report refers to transparency reports, we use that term here.

17 Illinois Senate Bill 315, Artificial Intelligence Safety Measures Act, Public Act 104-0538, § 10(d).

18 SB 813 Independent Verification Organizations, SB 813, California Legislature 2025–2026 Regular Session, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB813.

19 AB 1405 Artificial Intelligence: Auditors: Registration, AB 1405, California Legislature 2025–2026 Regular Session (2026), https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260AB1405.

20 Phillips-Robins and Singer, "With the RAISE Act, New York Aligns With California on Frontier AI Laws."

21 Illinois Senate Bill 315, Artificial Intelligence Safety Measures Act.

22 Phillips-Robins and Singer, "With the RAISE Act, New York Aligns With California on Frontier AI Laws."

23 Illinois Senate Bill 315, Artificial Intelligence Safety Measures Act.

24 Bommasani, Singer, et al., "The California Report on Frontier AI Policy"; Rishi Bommasani et al., "Advancing Science- and Evidence-Based AI Policy," *Science* 389, no. 6759 (July 31, 2025): 459–461, https://doi.org/10.1126/science.adu8449; and Stephen Casper, David Krueger, and Dylan Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy," arXiv, February 13, 2025, https://arxiv.org/abs/2502.09618.

25 The paper studies the AI in Government Act of 2020, Executive Order 13,859 on AI Leadership, and Executive Order 13,960 on AI in Government. See Christie Lawrence, Isaac Cui, and Daniel E. Ho, "Implementation Challenges to Three Pillars of America's AI Strategy," Stanford Institute for Human-Centered Artificial Intelligence and Stanford Regulation, Evaluation, and Governance Lab, December 2022, https://hai-production.s3.amazonaws.com/files/2022-12/HAIRegLab%20White%20Paper%20-%20Implementation%20Challenges%20to%20Three%20Pillars%20of%20America%E2%80%99s%20AI%20Strategy.pdf.

26 For the original text, see Jeffrey L. Pressman and Aaron Wildavsky, *Implementation: How Great Expectations in Washington Are Dashed in Oakland; Or, Why It's Amazing that Federal Programs Work at All, This Being a Saga of the Economic Development Administration as Told by Two Sympathetic Observers Who Seek to Build Morals on a Foundation* (University of California Press, 1979). For a more recent review, see Kai Wegrich, "Jeffrey L. Pressman and Aaron B. Wildavsky, Implementation," in The Oxford Handbook of Classics in Public Policy and Administration, ed. Steven J. Balla, Martin Lodge, and Edward C. Page (Oxford: Oxford University Press, 2016), https://doi.org/10.1093/oxfordhb/9780199646135.013.10.

27 Pressman and Wildavsky, *Implementation*, xxiii.

28 Pressman and Wildavsky, *Implementation*, xxii.

29 Pressman and Wildavsky, *Implementation*, 64–65.

30 Yoshua Bengio et al., "International AI Safety Report 2026," AI Security Institute, UK Department for Science, Innovation & Technology, February 2026, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026. For an in-depth discussion of this historical analogy, see Bommasani, Singer, et al., "The California Report on Frontier AI Policy," chapter 2. On tobacco, see, for example, David Michaels, Doubt Is Their Product: How Industry's Assault on Science Threatens Your Health (Oxford University Press, 2008); and Robert N. Proctor, Cancer Wars: How Politics Shapes What We Know and Don't Know About Cancer (Basic Books, 1994). On fossil fuels, see, for example, American Petroleum Institute, Two Energy Futures: A

National Choice for the '80s (American Petroleum Institute, 1980); and Geoffrey Supran, Stefan Rahmstorf, and Naomi Oreskes, "Assessing ExxonMobil's Global Warming Projections," Science 379, no. 6628 (2023): https://doi.org/10.1126/science.abk0063.

31 Pressman and Wildavsky, *Implementation*, 35.

32 Pressman and Wildavsky, *Implementation*, 93.

33 Pressman and Wildavsky, *Implementation*, xxiv.

34 Pressman and Wildavsky, *Implementation*, 107.

35 Remco Zwetsloot, "Strengthening the U.S. AI Workforce: A Policy and Research Agenda," Center for Security and Emerging Technology, September 2019, https://cset.georgetown.edu/publication/strengthening-the-u-s-ai-workforce; Matt Sheehan, "Have Top Chinese AI Researchers Stayed in the United States?," Emissary, Carnegie Endowment for International Peace, December 3, 2025, https://carnegieendowment.org/emissary/2025/12/china-ai-researchers-us-talent-pool; and Anka Reuel, Lisa Soder, Ben Bucknall, and Trond Arne Undheim, "Position: Technical Research and Talent Is Needed for Effective AI Governance," Proceedings of the 41st International Conference on Machine Learning, PMLR 235 (2024): 42543–42557, https://proceedings.mlr.press/v235/reuel24a.html.

36 Bommasani, Singer, et al., "The California Report on Frontier AI Policy," chapter 2.

37 Katie Siek and Jasmine Batten, "Taulbee Survey 2024: Annual Report," Computing Research Association, June 23, 2025, https://datavisualization.cra.org/TaulbeeSurvey/CRA_Taulbee_Survey_Report_2024.html.

38 Yiqin Fu and Alasdair Phillips-Robins, "When Should Congress Preempt State AI Law? The Lessons of Past Technologies," Carnegie Endowment for International Peace, September 9, 2025, https://carnegieendowment.org/research/2025/09/congress-preempt-state-ai-law-the-lessons-of-past-technologies.

39 Fu and Phillips-Robins, "When Should Congress Preempt State AI Law?"

40 U.S. Center for Advancing Innovation and Standards for Super Intelligence, National Institute of Standards and Technology, https://www.nist.gov/caissi.

41 Jonah Weinbaum and Arthur Tellis, "What Will it Cost for the US to Be Ready for the Next Big AI Breakthrough?," Institute for Progress, May 13, 2026, https://ifp.org/funding-for-caisi.

42 Weinbaum and Tellis, "What Will It Cost for the US to Be Ready for the Next Big AI Breakthrough?"; and "About," AI Security Institute, Government of the United Kingdom, https://www.aisi.gov.uk/about.

43 Fu and Phillips-Robins, "When Should Congress Preempt State AI Law?"

44 Peverill Squire, "Replication Data for: A Squire Index Update: Stability and Change in Legislative Professionalization, 1979-2021," UNC Dataverse, v1, https://doi.org/10.15139/S3/2VORBN.

45 "Talent Exchange Guide for State Agencies," California Department of Human Resources, accessed September 17, 2026, https://www.calhr.ca.gov/about-calhr/divisions-programs/workforce-development-division/workforce-planning/workforce-planning-toolkit/talent-exchange-guide-state-agencies.

46 This report's lead writer served as an adviser to the California Council on Science & Technology for its selection process for Scientific Advisor.

47 Cal. Gov. Code § 11181.

48 Cal. Gov. Code §§ 11180, 11181(e); Craib v. Bulmash, 49 Cal. 3d 475, 482 (1989).

49 Wayne B. Gray and Jay P. Shimshack, "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence," *Review of Environmental Economics and Policy* 5, no. 1 (2011), https://www.journals.uchicago.edu/doi/10.1093/reep/req017.

50 Cal. Bus. & Prof. Code § 22757.12(a)(8).

51 Cal. Bus. & Prof. Code § 22757.12(a)(5).

52 Cal. Bus. & Prof. Code § 22757.12(a)(1).

53 Cal. Bus. & Prof. Code § 22757.12(c).

54 Cal. Bus. & Prof. Code § 22757.12(b)(1)-(2).

55 Cal. Bus. & Prof. Code § 22757.12(c)(1).

56 Marta Ziosi et al., "Safety Frameworks and Standards: A Comparative Analysis to Advance Risk Management of Frontier AI," Oxford University Oxford Martin School AI Governance Initiative, October 9, 2025, https://aigi.ox.ac.uk/publications/safety-frameworks-and-standards-a-comparative-analysis-to-advance-risk-management-of-frontier-ai; Simeon Campos et al., "A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management," arXiv, February 10, 2025, https://arxiv.org/abs/2502.06656; Lily Stelling, Malcolm Murray, Simeon Campos, and Henry Papadatos, "Evaluating AI Companies' Frontier Safety Frameworks: Methodology and Results," arXiv, December 1, 2025, https://arxiv.org/abs/2512.01166; and Matteo Pistillo, "Towards Frontier Safety Policies Plus," arXiv, 2025, https://arxiv.org/abs/2501.16500.

57 OpenAI, "Preparedness Framework, Version 2," April 15, 2025, https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf; and OpenAI, "GPT-6 Astra System Card," September 3, 2026, https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf.

58 Anthropic, "Sharing Our Compliance Framework for California's Transparency in Frontier AI Act," Anthropic, December 19, 2025, https://www.anthropic.com/news/compliance-framework-SB53; Anthropic, "Responsible Scaling Policy, Version 3.1," effective April 2, 2026, https://www-cdn.anthropic.com/files/4zrzovbb/website/bf04581e4f329735fd90634f6a1962c13c0bd351.pdf; OpenAI, "Frontier Governance Framework," 2026, https://cdn.openai.com/pdf/e37d949b-8c9f-4d76-b99e-4272f4631a7e/openai-frontier-governance-framework.pdf; and OpenAI, "Preparedness Framework, Version 2."

59 Cal. Bus. & Prof. Code § 22757.12(c)(3).

60 Cal. Bus. & Prof. Code § 22757.12(c)(2).

61 Here, New York goes slightly further than California and New York: While SB 53 and Illinois's SB 315 both ask only that developers "describe" their "approach" to each specified element, RAISE requires them to set out "in detail" how they "handle" the same topics.

62 Cal. Bus. & Prof. Code § 22757.12(f)(1)-(2).

63 Cal. Bus. & Prof. Code § 22757.12(b)(2).

64 European Commission, "Commission Presents Template for General-Purpose AI Model Providers to Summarise the Data Used to Train Their Model," July 2025, https://digital-strategy.ec.europa.eu/en/news/commission-presents-template-general-purpose-ai-model-providers-summarise-data-used-train-their.

65 Frontier Model Forum, "Issue Brief: Components of Frontier AI Safety Frameworks," November 8, 2024, https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks.

66 For examples, see METR, "Example Safety and Security Framework (Draft)," updated November 26, 2025, https://metr.org/safety-security-framework.pdf; and International Organization for Standardization, "ISO/IEC WD TS 25568: Information Technology—Artificial Intelligence—Guidance on Addressing Risks in Generative AI Systems," https://www.iso.org/standard/88410.html. See also SaferAI, "An Overview of Existing and Potential Future GenAI/GPAI Standards," January 29, 2026, https://www.safer-ai.org/an-overview-of-existing-and-potential-future-genai-gpai-standards. The ISO standard is currently likely to be more than twelve months from publication.

67 Huw Roberts and Marta Ziosi, "Global AI Governance Through Technical Standards," Handbook on the Global Governance of Artificial Intelligence (forthcoming), August 2, 2025, https://doi.org/10.2139/ssrn.5376424.

68 European AI Office, "AI Act: Participate in the Drawing-Up of the First General-Purpose AI Code of Practice," European Commission, July 30, 2024, https://digital-strategy.ec.europa.eu/en/news/ai-act-participate-drawing-first-general-purpose-ai-code-practice.

69 Cal. Bus. & Prof. Code § 22757.12(a)(6).

70 Cal. Bus. & Prof. Code § 22757.12(b)(1)-(2).

71 See Cal. Bus. & Prof. Code § 22757.12(a).

72 See European AI Office, "General-Purpose AI Code of Practice," European Commission, July 10, 2025, https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. See especially Safety and Security chapter, measure 1.3(1)–(3).

73 For example, see the European Commission Guidelines for Providers of GPAI Models, which proposes quantitative thresholds for downstream modifiers at one-third of the FLOP threshold, 10^25, which in the EU AI Act defines a GPAI with systemic risk. See "Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models Under the AI Act," European Commission, July 18, 2025, https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act.

74 See Cal. Bus. & Prof. Code § 22757.13(a); see also European AI Office, "General-Purpose AI Code of Practice," Safety and Security chapter, measure 1.3(2).

75 Cal. Bus. & Prof. Code § 22757.12(a)(5).

76 Cal. Bus. & Prof. Code § 22757.12(b)(1)-(2).

77 Cal. Bus. & Prof. Code § 22757.12(c)(1)-(2).

78 Cal. Bus. & Prof. Code § 22757.11(d).

79 White House, Winning the Race: America's AI Action Plan, White House, July 2025, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

80 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), OJ L, 2024/1689, July 12, 2024, art. 55(1)(c), http://data.europa.eu/eli/reg/2024/1689/oj; and European AI Office, "General-Purpose AI Code of Practice," Safety and Security chapter.

81 Regulation (EU) 2024/1689, art. 73.

82 Cal. Bus. & Prof. Code § 22757.13(c)(1)-(2).

83 New York State Senate, Assembly Bill A9449, 2025–2026 Legislative Session, https://www.nysenate.gov/legislation/bills/2025/A9449; N.Y. Gen. Bus. Law § 1422; and Illinois Senate Bill 315, Artificial Intelligence Safety Measures Act.

84 Cal. Bus. & Prof. Code § 22757.13(h)-(j).

85 Kevin Wei and Lennart Heim, "Designing Incident Reporting Systems for Harms from General-Purpose AI," arXiv, November 8, 2025, https://arxiv.org/abs/2511.05914.

86 Cal. Bus. & Prof. Code § 22757.11(d).

87 Bengio et al., "International AI Safety Report 2026."

88 Cal. Bus. & Prof. Code § 22757.11(d); for an overview of current understanding of loss of control risks from deceptive behaviors, see Bengio et al., "International AI Safety Report 2026."

89 Bengio et al., "International AI Safety Report 2026."

90 This definition also excludes ordinary model errors. Hallucinations and confabulations are misleading, but they are not systematically directed at undermining the developer's understanding of the model or the effectiveness of its controls and monitoring.

91 Cal. Bus. & Prof. Code § 22757.11(c)(1)(A)–(C).

92 Cal. Bus. & Prof. Code § 22757.13(c)(1).

93 Cal. Bus. & Prof. Code § 22757.13(c)(2).

94 Norgart v. Upjohn Co., 21 Cal. 4th 383 (1999).

95 Cal. Bus. & Prof. Code § 22757.13(c)(3).

96 Cal. Bus. & Prof. Code § 22757.13(e).

97 Cal. Bus. & Prof. Code § 22757.13(f).

98 CISA's relationship with Carnegie Mellon is managed by the CERT Coordination Center (CERT/CC), which sits within the university's Software Engineering Institute, a federally funded research and development center. CERT/CC was established at DARPA's direction in 1988 and now performs coordinated vulnerability disclosure work sponsored by CISA: It builds tools for secure disclosure, receives vulnerability reports, coordinates analysis and remediation among affected vendors, and publishes public Vulnerability Notes. See CERT Division, Software Engineering Institute, Carnegie Mellon University, accessed September 21, 2026, https://www.sei.cmu.edu/divisions/cert.

99 Wei and Heim, "Designing Incident Reporting Systems."

100 49 U.S.C. § 1154(b).

101 Cal. Bus. & Prof. Code § 22757.15(b).

102 Cal. Bus. & Prof. Code § 22757.13(e)(2).

103 Cal. Bus. & Prof. Code § 22757.13(e)(2).

104 Cal. Bus. & Prof. Code § 22757.13(c)(4).

105 OpenAI, "Periodic Summary of Internal Use Catastrophic-Risk Assessments," September 2026, https://cdn.openai.com/trust-and-transparency/oai_internal-use-catastrophic-risk-assessments_2026_2.pdf.

106 See Robert M. Bowen, Andrew C. Call, and Shivaram Rajgopal, "Whistle-Blowing: Target Firm Characteristics and Economic Consequences," AAA 2007 Financial Accounting & Reporting Section (FARS) Meeting Papers, December 26, 2009, http://dx.doi.org/10.2139/ssrn.890750.

107 California Labor Code § 1107.1(a); and Artificial Intelligence Safety Measures Act, Pub. Act 104-0538, § 20(a) (Ill. 2026).

108 Directive (EU) 2019/1937.

109 See, for example, 5 U.S.C. § 2302 (public employees); 18 U.S.C. § 1514A (private employees in fraud cases).

110 Article 87 of Regulation (EU) 2024/1689, and California Labor Code § 1107.

111 Connecticut Public Act No. 26-15 (2026), Act Concerning Online Safety, https://www.cga.ct.gov/2026/act/pa/pdf/2026PA-00015-R00SB-00005-PA.pdf.

112 California Labor Code § 1102.5, upon which SB 53's whistleblower protection provision is substantially modeled, forbids employers from preventing or retaliating against employees for disclosing information "if the employee has a reasonable cause to believe that the information discloses a violation of state or federal statute, or a violation of or noncompliance with a local, state, or federal rule or regulation." See also NY Labor Law § 740.

113 I. J. Alexander Dyck, Adair Morse, and Luigi Zingales, "Who Blows the Whistle on Corporate Fraud?," AFA 2007 Chicago Meetings Paper, 1st Annual Conference on Empirical Legal Studies, Forthcoming, ECGI - Finance Working Paper No. 156/2007, October 1, 2008, http://dx.doi.org/10.2139/ssrn.891482.

114 Dyck, Morse, and Zingales, "Who Blows the Whistle on Corporate Fraud?" One significant exception, in which whistleblowers may receive tangible benefit, is whistleblowing under the federal False Claims Act, which

authorizes private citizens to recover damages for fraud perpetrated against the government. While SB 53 allows successful parties to recoup attorneys' fees, it offers no comparable incentive for disclosure.

115 Bowen, Call, and Rajgopal, "Whistle-Blowing: Target Firm Characteristics and Economic Consequences."

116 See People ex rel. Garcia-Brower v. Kolla's, Inc. (2023) 14 Cal.5th 719.

117 By supporting companies' ability to exercise reasonable care, it may also help them avoid tort liability later on. See, for example, Ketan Ramakrishnan, "Tort Law is the Best Way to Regulate AI," *Wall Street Journal*, September 24, 2024, https://www.wsj.com/opinion/tort-law-is-the-best-way-to-regulate-ai-california-legal-liability-065e1220?msockid=3c8f3b4afc606d3b2dba2ff5fd676c4d.

118 California Labor Code § 1102.5; NY Labor Law § 740.

119 See draft text at California Senate Bill No. 53, "Bill Text: CA SB53 | 2025-2026 | Regular Session | Amended," LegiScan, https://legiscan.com/CA/text/SB53/id/3268028.

120 These concepts occur elsewhere in whistleblower protection law. See, for example, California Labor Code § 1102.5 (substantially identical language regarding "reasonable cause to believe"). See also the federal Whistleblower Protection Act: "'Disclosure' means a formal or informal communication or transmission, but does not include a communication concerning policy decisions that lawfully exercise discretionary authority unless the employee or applicant providing the disclosure reasonably believes that the disclosure evidences—(i) any violation of any law, rule, or regulation; or (ii) gross mismanagement, a gross waste of funds, an abuse of authority, or a substantial and specific danger to public health or safety." See 5 U.S.C. § 2302(a)(2)(D).

121 People ex rel. Garcia-Brower v. Kolla's, Inc., (2023) 14 Cal.5th 719.

122 Italics in the original. See Killgore v. SpecPro Professional Services, LLC (9th Cir. 2022) 51 F.4th 973, 988, quoted in Contreras v. Green Thumb Produce (Cal. Ct. App. 2025).

123 Ross v. County of Riverside (2019) 36 Cal.App.5th 580, 592.

124 Love v. Motion Industries, Inc. (N.D. Cal. 2004) 309 F.Supp.2d 1128, 1135.

125 See 42 U.S.C. § 5851; and U.S. Department of Labor, Occupational Safety and Health Administration Fact Sheet, "Whistleblower Protection for Nuclear Industry Workers," April 2018, https://www.osha.gov/sites/default/files/publications/OSHA3948.pdf.

126 See Congressional Research Service, Report R466, "Federal Scientific Integrity Policies: A Primer," November 9, 2022, https://www.congress.gov/crs-product/R46614.

127 People ex rel. Garcia-Brower v. Kolla's, Inc., (2023) 14 Cal.5th 719.

128 See California Labor Code § 1102.5; Whistleblower Protection Act 5 U.S.C. § 2302(a)(2)(D). See also California Labor Code § 1102.5; and NY Labor Law § 740. (New York protects disclosures of "a substantial and specific danger to public health or safety.")

129 Chambers v. Department of the Interior, 515 F.3d 1362, 1369 (Fed. Cir. 2008).

130 Chambers v. Department of the Interior, 515 F.3d 1362, 1369 (Fed. Cir. 2008).

131 Chambers v. Department of the Interior, 515 F.3d 1362, 1369 (Fed. Cir. 2008).

132 Connecticut Public Act No. 26-15 (2026).

133 Large frontier developers are subject to additional procedural obligations, to support transparent and diligent handling of complaints. These developers must also establish internal processes to enable anonymous reporting, provide quarterly reports to officers and directors, and provide monthly updates to the disclosing employee "regarding the status of the large frontier developer's investigation of the disclosure and the actions taken . . . in response." California Labor Code § 1107.1(e).

134 California Senate Bill 53, *Transparency in Frontier Artificial Intelligence Act*.

135 In addition to the financial and operational impacts identified in Bowen, Call, and Rajgopal, "Whistle-Blowing: Target Firm Characteristics and Economic Consequences," companies have numerous reasons to prefer internal reporting, including to prevent leakage of sensitive information (for example, intellectual property or confidential business information), to avoid reputational harm, and, importantly, to enable themselves to consider and act on the legitimate concerns of knowledgeable insiders.

136 California Labor Code § 1107.1(e)(1).

137 California Labor Code § 1107.1(e)(1).

138 Miles Brundage et al., "Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies," arXiv:2601.11699, preprint, arXiv, February 7, 2026, https://doi.org/10.48550/arXiv.2601.11699.

139 For various challenges associated with evaluations, see generally Lily Stelling et al., "The Promises and Limitations of Evaluation-Based AI Governance: Evidence from Expert Interviews," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 9, no. 4 (2026): 2421–2434, https://doi.org/10.1609/aies.v9i4.43317; and Thomas Liao et al., "Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning," *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1 (December 2021), https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html.

140 Anthropic, "Partnering with Accenture on Embedded Evaluation," Anthropic, September 18, 2026, https://perma.cc/9CJS-96PG; Sam Altman (@sama), "I agree with Dario that we need to pace the frontier. This has been a primary topic of discussions we've had at OpenAI in recent weeks. Committing to having independent evaluators with employee-like access is a great idea, and we will do the same. We'll have more to share soon.," X post, September 12, 2026, https://perma.cc/N39N-M9AZ; and Elon Musk (@elonmusk), "SpaceX will probably do the same," X, October 16, 2026, https://perma.cc/N39N-M9AZ.

141 Gregory C. Allen and Georgia Adamson, "The AI Safety Institute International Network: Next Steps and Recommendations," Center for Strategic and International Studies, October 30, 2024, https://perma.cc/3955-4ZP3.

142 METR, "Common Elements of Frontier AI Safety Policies," METR, December 16, 2025, https://perma.cc/F689-Y8XZ; and Redwood Research, https://www.redwoodresearch.org.

143 California Senate Bill 53, Transparency in Frontier Artificial Intelligence Act.

144 AB 1405 Artificial Intelligence: Auditors: Registration.

145 SB 813 Independent Verification Organizations.

146 SB 813 Independent Verification Organizations, § 8898.4(a)(3).

147 Brundage et al., "Frontier AI Auditing."

148 A benchmark is a standardized test administered to AI systems, that is, a benchmark presents AI systems with a fixed set of questions to answer. Benchmarks can be agentic or non-agentic. Agentic benchmarks ask AI systems to actually perform a task while interacting with external tools (for example, an email tool to send or receive emails), while non-agentic benchmarks only ask AI systems to provide a textual or multimodal (for example, image or video) response without interacting with external tools. A task in an agentic benchmark might ask an AI system to hack into a cybersecurity system; a question in a non-agentic benchmark might ask an AI system to provide written instructions for conducting such a hack rather than actually carrying out the hack autonomously. See Anka Reuel et al., "BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices," *Advances in Neural Information Processing Systems* 37 (December 2024): 21763–813, https://doi.org/10.52202/079017-0685.

149 A red-teaming evaluation asks human experts, large language models, or some combination of both to interact with an evaluated AI system to test whether it is secure against adversarial attacks. As opposed to

the fixed question banks used in benchmarks, red-teaming is dynamic, flexible, and seeks to identify system vulnerabilities, new behaviors, or unknown risks. See Michael Feffer et al., "Red-Teaming for Generative AI: Silver Bullet or Security Theater?," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (2024): 421–37, https://doi.org/10.1609/aies.v7i1.31647.

150 An uplift study aims to measure the level of assistance an AI system gives to a human user in completing a task, often using randomized controlled trial methodology. Because access to AI systems usually increases task performance, these evaluations are called uplift studies. One domain in which uplift studies have been used is in evaluations of biosecurity risk. See Patricia Paskov et al., "RCTs & Human Uplift Studies: Methodological Challenges and Practical Solutions for Frontier AI Evaluation," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 9, no. 3 (2026): 2036–52, https://doi.org/10.1609/aies.v9i3.43287.

151 Many different methods exist for evaluating different properties of AI systems, and which specific method is appropriate depends significantly on risk vectors and the properties being measured. For instance, uplift studies are often used to measure the potential for AI to be misused by malicious actors to create bioweapons, but they are not used much, if at all, for alignment evaluations that intend to measure character traits in AI systems such as honesty or sycophancy.

152 METR, "Common Elements of Frontier AI Safety Policies."

153 The White House (@WhiteHouse), "White House Accord on Super Intelligence," X post, September 30, 2026, 9:44 a.m., https://perma.cc/4FWA-QZ5F; and Sara Dorn, "White House Releases 'Accord' Between Billionaire AI Execs: Here's What It Says," *Forbes*, September 30, 2026, https://www.forbes.com/sites/saradorn/2026/09/30/white-house-releases-accord-between-billionaire-ai-execs-heres-what-it-says.

154 Jacob Charnock et al., "Embedded Assessments for Frontier AI," arXiv:2609.25413, preprint, arXiv, September 21, 2026, https://doi.org/10.48550/arXiv.2609.25413; and Apollo Research, "Embedded Evaluators Are Necessary for Meaningful External Testing," *Apollo Research*, September 29, 2026, https://perma.cc/8DAS-ATAS.

155 Rebecca Bellan, "Anthropic and OpenAI Want to Embed Safety Evaluators. Will They Really Be Independent?," *TechCrunch*, September 16, 2026, https://perma.cc/JKT7-2GGG.

156 "An Update on Independent Evaluators," *Anthropic*, October 15, 2026, https://perma.cc/MGV7-BZUS; Anthropic, "Partnering with Accenture on Embedded Evaluation."

157 David Rein, "Red-Teaming Anthropic's Internal Agent Monitoring Systems," METR, March 26, 2026, https://perma.cc/N4PW-2PCN; Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk, "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," METR, August 26, 2026, https://perma.cc/4S7B-MBPX.

158 AB 1405 Artificial Intelligence: Auditors: Registration, § 11549.80(d).

159 The deadline specified in AB 1405 is January 1, 2029, but California Governor Gavin Newsom's executive order moved the deadline up to December 1, 2027. "Executive Order N-9-26," State of California Executive Department, September 18, 2026, https://perma.cc/5E5F-43KS.

160 The deadline specified in SB 813 is January 1, 2028, but Newsom's executive order moved the deadline up to May 1, 2027. "Executive Order N-9-26," State of California Executive Department.

161 SB 813 Independent Verification Organizations, § 8898(e).

162 Olawale Salaudeen et al., "Measurement to Meaning: A Validity-Centered Framework for AI Evaluation," arXiv:2505.10573, preprint, arXiv, revised June 26, 2025, https://doi.org/10.48550/arXiv.2505.10573.

163 OpenAI, "Separating Signal from Noise in Coding Evaluations," OpenAI, July 8, 2026, https://openai.com/index/separating-signal-from-noise-coding-evaluations.

164 Ali Ansari et al., "How Good Are Frontier Models at Physics? Expert Re-Grading Reveals Broken Evaluations and Near-Saturation of Leading Benchmarks," arXiv:2609.13009, preprint, arXiv, September 11, 2026, https://doi.org/10.48550/arXiv.2609.13009.

165 See generally, Evi Fuelle and Ryan Pan, "AI Is Sprinting Past Human Control. Can Testing and Evaluation Catch Up?," Atlantic Council, September 21, 2026, https://perma.cc/WB5H-KUKS. For more technical details, see Maia Hamin and Benjamin Edelman, "Cheating on AI Agent Evaluations," U.S. Center for AI Standards and Innovation, November 28, 2025, https://perma.cc/E4GH-TX5C; Hao Wang et al., "Do Androids Dream of Breaking the Game? Systematically Auditing AI Agent Benchmarks with BenchJack," arXiv:2605.12673, preprint, arXiv, May 12, 2026, https://doi.org/10.48550/arXiv.2605.12673; and Joe Needham et al., "Large Language Models Often Know When They Are Being Evaluated," arXiv:2505.23836, version 3, preprint, arXiv, revised July 16, 2025, https://doi.org/10.48550/arXiv.2505.23836.

166 Andrew M. Bean et al., "Measuring What Matters: Construct Validity in Large Language Model Benchmarks," *Advances in Neural Information Processing Systems* 38 (April 2026), https://doi.org/10.52202/085713-0590.

167 See Elizabeth Seger and Darcy Ward, "AI-Assurance Ecosystems: Building the Infrastructure to Enable Confident Adoption of Safe AI," Tony Blair Institute for Global Change, September 18, 2026, https://institute.global/insights/tech-and-digitalisation/ai-assurance-ecosystems-building-infrastructure-to-enable-confident-adoption-of-safe-ai; and Brundage et al., "Frontier AI Auditing."

168 Accenture, "Accenture and Anthropic Launch Multi-Year Partnership to Drive Enterprise AI Innovation and Value Across Industries," Accenture, December 9, 2025, https://perma.cc/2J3W-NX22.

169 Stephen Casper et al., "Black-Box Access Is Insufficient for Rigorous AI Audits," *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY), FAccT '24, June 5, 2024, 2254–72, https://doi.org/10.1145/3630106.3659037.

170 OpenAI, "OpenAI GPT-4.5 System Card," OpenAI, February 27, 2025, https://perma.cc/YJ4U-DUY9.

171 Xiangyu Qi et al., "AI Risk Management Should Incorporate Both Safety and Security," arXiv:2405.19524, preprint, arXiv, May 29, 2024, https://doi.org/10.48550/arXiv.2405.19524.

172 Alexander Wan et al., "The 2025 Foundation Model Transparency Index," arXiv:2512.10169, preprint, arXiv, December 11, 2025, https://doi.org/10.48550/arXiv.2512.10169; and Leon Staufer et al., "The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems," *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, June 25, 2026, 1536–76, https://doi.org/10.1145/3805689.3806728.

173 SpaceXAI and Cursor, "Model Card Grok 4.5," July 14, 2026, https://perma.cc/V9AD-QJUZ.

174 See Meera Desai et al., "What AI Benchmarks Actually Measure: Adapting Convergent and Discriminant Validity to Interrogate Fifty-Six AI Benchmarks," paper presented at Third Conference on Language Modeling, August 24, 2026, https://openreview.net/forum?id=889XnQKyhM.

175 Cristian Trout et al., "Underwriting the Agent Economy: The Blueprint for an AI Insurance Stack," arXiv:2607.11999, preprint, arXiv, July 15, 2026, https://doi.org/10.48550/arXiv.2607.11999.

176 "ML Engineer Salary 2026: Tier Benchmarks and Total Comp," MLEngineerSalary.Com, April 28, 2026, https://perma.cc/2H2T-GT2B.

177 Zoë Schiffer, "Here's What Mark Zuckerberg Is Offering Top AI Talent," *Wired*, July 1, 2025, https://perma.cc/EPZ5-J7NP.

178 "Machine Learning Research Scientist, Evaluations," Scale AI, September 2026, https://perma.cc/28DT-UKNJ. Scale AI was operating at the time of this writing in September 2026.

179 SB 813 Independent Verification Organizations, § 8898.3(a).

180 SB 813 Independent Verification Organizations, § 8898.1(b).

181 AB 1405 Artificial Intelligence: Auditors: Registration, § 11549.82(a)(3).

182 California Senate Bill 53, Transparency in Frontier Artificial Intelligence Act, § 1107.1; and AB 1405 Artificial Intelligence: Auditors: Registration, § 11549.83(g)(1).

183 Peter Wills, "Regulatory Supervision of Frontier AI Developers," SSRN Scholarly Paper no. 5122871 (Social Science Research Network, February 2, 2026), https://doi.org/10.2139/ssrn.5122871.

184 SB 813 Independent Verification Organizations, § 8898.1(c)(1).

185 "International Evaluation Best Practice and Open Questions in AI Measurement," UK AI Security Institute, July 23, 2026, https://perma.cc/GFQ4-8TSA; U.S. Center for AI Standards and Innovation, "NIST AI 800-2 (Initial Public Draft): Towards Best Practices for Automated Benchmark Evaluations," January 30, 2026, https://doi.org/10.6028/NIST.AI.800-2.ipd; Yuxuan Zhu et al. "Establishing Best Practices in Building Rigorous Agentic Benchmarks," *Advances in Neural Information Processing Systems* 38 (April 2026), https://doi.org/10.52202/085713-5547; Reuel et al., "BetterBench"; Guidelight AI Standards, "AI Control: An Assessment of Frontier Practices," Guidelight AI Standards, August 25, 2026, https://perma.cc/R7HV-YTVC; Patricia Paskov et al., "Preliminary Suggestions for Rigorous GPAI Model Evaluations," RAND, May 1, 2025, https://doi.org/10.7249/PEA3971-1; and Paskov et al., "RCTs for Frontier AI Governance."

186 SB 813 Independent Verification Organizations, § 8898.1(b)(6).

187 OpenMined, "Secure Enclaves for AI Evaluation," OpenMined, December 12, 2024, https://perma.cc/CZU7-SCYC; and Andrew Trask et al., *Double Blind Evals: Resolving the Dual Confidentiality Dilemma in AI Safety Auditing* (AVERI and Google DeepMind, 2026), https://perma.cc/9AV4-NFTN.

188 Greenblatt et al., "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident."

189 AB 1405 Artificial Intelligence: Auditors: Registration, § 11549.83(f)(1).

190 SB 813 Independent Verification Organizations, § 8898.1(b)(3) and § 8898.1(b)(5).

191 See AB 1405 Artificial Intelligence: Auditors: Registration, § 11549.83(f)(1).

192 Conrad Stosz et al., "AEF-1: Minimum Operating Conditions for Independent Third Party AI Evaluations," AI Evaluators Forum, December 2025, https://perma.cc/J2XT-HJYW.

193 Epoch AI, for instance, has recently introduced "benchmark reviews" that audit benchmarks for quality issues. Epoch AI, "Benchmarks," Epoch AI, accessed September 25, 2026, https://perma.cc/3UTX-MAAS.

194 OpenAI, "Separating Signal from Noise in Coding Evaluations."

195 For a longer (but still incomplete) list of possible issues with AI evaluations, see Ayrton San Joaquin, Rokas Gipiškis, Leon Staufer, and Ariel Gil, "Deprecating Benchmarks: Criteria and Framework," arXiv:2507.06434, preprint, arXiv, July 8, 2025, https://doi.org/10.48550/arXiv.2507.06434.

196 SB 813 Independent Verification Organizations, § 8898.1(c)(2)(B).

197 AB 1405 Artificial Intelligence: Auditors: Registration, § 11549.83(f)(1)(F).

198 Avijit Ghosh et al., "Evaluation Cards: An Interpretive Layer for AI Evaluation Reporting," arXiv:2606.09809, preprint, arXiv, revised June 9, 2026, https://doi.org/10.48550/arXiv.2606.09809; and Jan Batzner et al., "Every Eval Ever: A Unifying Schema and Community Repository for AI Evaluation Results," arXiv:2606.14516, preprint, arXiv, June 12, 2026, https://doi.org/10.48550/arXiv.2606.14516.

199 California Senate Bill 53, Transparency in Frontier Artificial Intelligence Act, § 22757.12(f)(1).

200 Keleigh Friedrich, "CCST Launches California AI Science Residency Program on Frontier AI Safety with Appointment of First AI Science Advisors to the California Governor's Office of Emergency Services and Department of Technology," *California Council on Science & Technology (CCST)*, July 7, 2026, https://perma.cc/2ZWS-ZHR7.

201 Aidan Homewood et al., "Third-Party Compliance Reviews for Frontier AI Safety Frameworks," arXiv:2505.01643, preprint, arXiv, revised July 4, 2025, https://doi.org/10.48550/arXiv.2505.01643.

202 California Senate Bill 53, Transparency in Frontier Artificial Intelligence Act, § 22757.12(a).

203 Illinois Senate Bill 315, *Artificial Intelligence Safety Measures Act*, § 10(d).

204 SB 813 Independent Verification Organizations, § 8898.4(a)(3).

205 OpenAI, "GPT-6 Astra System Card."

206 METR, "Our Approach to Working with AI Companies," METR, October 14, 2026, https://perma.cc/BG6W-6YYS; Transluce, "Template Memorandum of Understanding Between Transluce and Anthropic, OpenAI, and Google DeepMind, Respectively," Transluce, 2026, https://perma.cc/S3CD-8DPT; and Apollo Research, "Principles for Embedded Evaluations," Apollo Research, September 30, 2026, https://perma.cc/6NJQ-VK38.

207 Brundage et al., "Frontier AI Auditing."

208 For instance, AI developers have reported evaluations run by for-profit evaluators including Gray Swan, Irregular, and Faculty AI. Sam Link, "Gray Swan, The AI Security Company Trusted by Every Major Frontier Lab, Raises $40M Series A," Gray Swan, May 28, 2026, https://perma.cc/5WVY-CRY9; Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4," Anthropic, May 2025, https://perma.cc/S2E7-9VHR; and Aaditya Singh et al., "OpenAI GPT-5 System Card," arXiv:2601.03267, preprint, arXiv, December 19, 2025, https://doi.org/10.48550/arXiv.2601.03267.

209 Inioluwa Deborah Raji et al., "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance," *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA), AIES '22, July 27, 2022, 557–71, https://doi.org/10.1145/3514094.3534181.

210 For emerging trends and open questions in agentic AI evaluation, see Felix Chen et al., "Emerging Trends in AI Agent Evaluation," preprint, October 16, 2026, https://perma.cc/3LAJ-QFMA.

211 See Konstantinos Voudouris et al., "AI Alignment Is a Human Problem," preprint, PsyArXiv, revised May 30, 2026, https://osf.io/preprints/psyarxiv/zqngj_v1/; Joshua Levy et al., "AI Persuasion as a Threat to Human Control," arXiv:2609.14796, preprint, arXiv, revised September 16, 2026, https://doi.org/10.48550/arXiv.2609.14796; Jeremy Qin et al., "LLM Agents Can Easily Tamper With Their Own Traces," arXiv:2609.30266, preprint, arXiv, September 24, 2026, https://doi.org/10.48550/arXiv.2609.30266; and Jonas C. Ditz et al., "Secure Human Oversight of AI: Threat Modeling in a Socio-Technical Context," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 9, no. 2 (2026): 843–56, https://doi.org/10.1609/aies.v9i2.43187.

212 Illinois Senate Bill 315, *Artificial Intelligence Safety Measures Act*, § 10(a).

213 METR, "Frontier Risk Report (February to March 2026)," May 19, 2026, https://metr.org/blog/2026-05-19-frontier-risk-report.

214 In general, though perhaps not in all cases, an AI developer's internally deployed models are at least as capable as its publicly deployed models.

215 Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Joe Kwon and Stephen Casper, "Internal Deployment Gaps in AI Regulation," arXiv, January 12, 2026, https://arxiv.org/abs/2601.08005.

216 California Senate Bill 53, *Transparency in Frontier Artificial Intelligence Act*. See also Cal. Bus. & Prof. Code § 22757.12(a)(10) and § 22757.12(d).

217 New York State Legislature, "Responsible AI Safety and Education Act"; and S.B. 315, 104th Gen. Assemb., Reg. Sess. (Ill. 2026).

218 Compare N.Y. Gen. Bus. Law § 1421(1)(j) and § 1422(2) with Cal. Bus. & Prof. Code § 22757.12(a)(10) and § 22757.12(d). The internal use language is largely identical.

219 N.Y. Gen. Bus. Law §1422(2), as added by S.8828, ch. 96 (2026) (schedule "agreed upon by the office"); Artificial Intelligence Safety Measures Act §10(e) (Ill. Pub. Act 104-0538) (schedule "agreed upon by the Agency"). Compare Cal. Bus. & Prof. Code §22757.12(d), which permits an alternative schedule "specified by the large frontier developer and communicated in writing," without agency agreement.

220 "Frontier Risk Report," METR.

221 Also in May 2026, CAISSI announced that it had "completed more than 40 such evaluations, including on state-of-the-art models that remain unreleased." This announcement was briefly available for several days before being taken down for unclear reasons. Executive Order 14409 (June 2, 2026) subsequently directed the secretary of the treasury and the directors of the National Security Agency and CISA to develop and maintain a classified benchmarking process for the cyber capabilities of advanced AI models. CAISSI was separately reported to have been directed to stop publishing its evaluation findings. See U.S. Center for Advancing Innovation and Standards for Super Intelligence, "CAISI Signs Agreements Regarding Frontier AI National Security Testing With Google DeepMind, Microsoft and xAI," U.S. National Institute of Standards and Technology, May 5, 2026, archived at https://web.archive.org/web/20260505134747/https://www.nist.gov/news-events/news/2026/05/caisi-signs-agreements-regarding-frontier-ai-national-security-testing; and Letter from U.S. Senator Ted Budd to the National Cyber Director and the Director of the Office of Science and Technology Policy Regarding CAISI, June 26, 2026, https://www.budd.senate.gov/wp-content/uploads/2026/06/CAISI-Letter_.pdf.

222 Nikola Jurkovic et al., "Review of the Anthropic Summer 2025 Pilot Sabotage Risk Report," METR, October 28, 2025, https://metr.org/blog/2025-10-28-sabotage-report-review; and Nikola Jurkovic et al., "Review of the Anthropic Sabotage Risk Report: Claude Opus 4.6," METR, March 12, 2026, https://metr.org/blog/2026-03-12-sabotage-risk-report-opus-4-6-review.

223 European Commission, "Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models Under the AI Act."

224 European Commission, "Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models Under the AI Act."

225 European Commission, "Guidelines on GPAI Obligations"; European AI Office, "General-Purpose AI Code of Practice," Safety and Security chapter; and Oscar Delaney et al., "Risk Reporting for Developers' Internal AI Model Use," Institute for AI Policy and Strategy, April 29, 2026, https://www.iaps.ai/research/risk-reporting-for-developers-internal-ai-model-use.

226 European Commission, "Guidelines on GPAI Obligations"; European AI Office, "General-Purpose AI Code of Practice"; Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Kwon and Casper, "Internal Deployment Gaps in AI Regulation."

227 European Commission, "Guidelines on GPAI Obligations"; European AI Office, "General-Purpose AI Code of Practice"; Matteo Pistillo, "Internal Deployment in the EU AI Act," arXiv, December 5, 2025, https://arxiv.org/abs/2512.05742; Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Kwon and Casper, "Internal Deployment Gaps in AI Regulation."

228 "Frontier Risk Report," METR.

229 OpenAI, "GPT-4," March 14, 2023, https://openai.com/index/gpt-4-research/.

230 Anthropic, "System Card: Claude Mythos Preview," April 7, 2026, https://www-cdn.anthropic.com/8b8380204f74670be75e81c820ca8dda846ab289.pdf; and Anthropic, "Alignment Risk Update: Claude Mythos Preview," April 10, 2026, https://anthropic.com/claude-mythos-preview-risk-report. Because Anthropic made Claude Mythos Preview available to outside partners, for a purpose besides developing or evaluating the model, the model is treated as deployed under SB 53. A public transparency report (or system

card) is required at deployment, and it must summarize the catastrophic risk assessments conducted under the developer's framework. Moreover, the framework must itself address risk from internal use, so internal use risk falls within the required public summary as well as the separate confidential quarterly filings to OES. Anthropic exceeded the statute's requirements, however, by providing specific internal use incidents and failure pathways that SB 53 requires only in summary form, if at all.

231 Anthropic Frontier Red Team, "Assessing Claude Mythos Preview's Cybersecurity Capabilities," April 7, 2026, https://www.anthropic.com/research/mythos-preview.

232 Anthropic Institute, "When AI Builds Itself," Anthropic, June 2026, https://www.anthropic.com/institute/recursive-self-improvement; and OpenAI, "Research Acceleration: The View Inside OpenAI," September 6, 2026, https://openai.com/index/research-acceleration-view-inside-openai.

233 "Frontier Risk Report," METR.

234 Anthropic, "Alignment Risk Update: Claude Mythos Preview"; and Anthropic, "System Card: Claude Mythos Preview."

235 Ashwin Acharya and Oscar Delaney, "Managing Risks from Internal AI Systems," Institute for AI Policy and Strategy, July 21, 2025, https://www.iaps.ai/research/managing-risks-from-internal-ai-systems.

236 Anthropic's Responsible Scaling Policy, Version 3.4 states that its risk reports will "cover internally deployed models when [it] determine[s] that these models could pose significant risks above and beyond those posed by our public models." See Anthropic, "Responsible Scaling Policy, Version 3.4." OpenAI's Preparedness Framework Version 2 applies to "any agentic system (including significant agents deployed only internally) that represents a substantial increase in the capability frontier." See OpenAI, "Preparedness Framework, Version 2."

237 OpenAI, "OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation," July 21, 2026, https://openai.com/index/hugging-face-model-evaluation-security-incident/

238 OpenAI, "OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation."

239 Anthropic reports that the earliest incidents date to April, that it notified the three affected organizations on Monday, July 27, and that "[t]he two organizations we were able to reach had not previously detected the activity or contacted us." No date within April is given though. See Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations," July 30, 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals

240 Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations."

241 Anthropic, "Responsible Scaling Policy, Version 3.4"; and OpenAI, "Preparedness Framework, Version 2."

242 Acharya and Delaney, "Managing Risks from Internal AI Systems"; and Charlotte Stix et al., "AI Behind Closed Doors: A Primer on the Governance of Internal Deployment," Apollo Research, April 17, 2025, https://www.apolloresearch.ai/governance/ai-behind-closed-doors-a-primer-on-the-governance-of-internal-deployment.

243 Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Kwon and Casper, "Internal Deployment Gaps in AI Regulation," sec. 2.1.

244 Stix et al., "AI Behind Closed Doors"; Acharya and Delaney, "Managing Risks from Internal AI Systems"; and Delaney, "AI Developers Should Report on Internal Model Use."

245 Stix et al., "AI Behind Closed Doors"; Acharya and Delaney, "Managing Risks from Internal AI Systems."

246 Ryan Greenblatt et al., "Alignment Faking in Large Language Models," arXiv, December 18, 2024, https://arxiv.org/abs/2412.14093.

247 Alexander Meinke et al., "Frontier Models Are Capable of In-Context Scheming," arXiv, December 6, 2024, https://arxiv.org/abs/2412.04984.

248 Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn, "Large Language Models Often Know When They Are Being Evaluated," arXiv, May 28, 2025, https://arxiv.org/abs/2505.23836.

249 Anthropic, "Risk Report: February 2026," 2026, 100–101, https://www-cdn.anthropic.com/08eca2757081e850ed2ad490e5253e940240ca4f.pdf; and Sella Nevo et al., "Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models," RAND Corporation, May 30, 2024, https://www.rand.org/pubs/research_reports/RRA2849-1.html.

250 "Frontier Risk Report," METR.

251 Anthropic, "System Card: Claude Mythos Preview," 61.

252 "Frontier Risk Report," METR.

253 California Senate Bill 53, *Transparency in Frontier Artificial Intelligence Act*.

254 Cal. Bus. & Prof. Code §22757.13(f).

255 Cal. Bus. & Prof. Code § 22757.12(a)(10).

256 Cal. Bus. & Prof. Code § 22757.12(a)(4).

257 Cal. Bus. & Prof. Code § 22757.13(f).

258 Delaney et al., "Risk Reporting for Developers' Internal AI Model Use."

259 Cal. Bus. & Prof. Code §22757.12(e). A statement made in violation of this subdivision is subject to a civil penalty of up to $1,000,000 per violation, dependent upon the severity of the violation and recoverable only in a civil action brought by the Attorney General. §22757.15(a)–(b).

260 Delaney et al., "Risk Reporting for Developers' Internal AI Model Use."

261 A benchmark is "saturated" when frontier models score at or near the maximum, leaving little to no headroom to detect further capability improvements.

262 Anthropic, "System Card: Claude Mythos Preview."

263 Anthropic, "Alignment Risk Update: Claude Mythos Preview."

264 Charnock et al., "What Should Frontier AI Developers Disclose About Internal Deployments?"

265 OpenAI: the deployment safeguards "were intentionally not enabled during this evaluation because it was aimed at testing cyber vulnerabilities." Anthropic: the models "ran without the standard safeguards we deploy when we make the model generally available," though they retained their model-specific safety training and were not "helpful-only" variants. See Anthropic, "Investigating Three Real-World Incidents"; and OpenAI, "OpenAI and Hugging Face Partner."

266 Cal. Bus. & Prof. Code § 22757.12(d).

267 N.Y. Gen. Bus. Law § 1422(2)(a), as added by S.8828, ch. 96 (2026), https://legislation.nysenate.gov/pdf/bills/2025/S8828.

268 Stix et al., "AI Behind Closed Doors."

269 Anthropic, "Alignment Risk Update: Claude Mythos Preview."

270 Acharya and Delaney, "Managing Risks from Internal AI Systems."

271 Stix et al., "AI Behind Closed Doors."

272 Pistillo, Matteo. 2025. Internal Deployment in the EU AI Act. https://arxiv.org/abs/2512.05742.

273 Kwon and Casper, "Internal Deployment Gaps in AI Regulation."

274 Cal. Bus. & Prof. Code § 22757.12(a)(4).

275 "Frontier Risk Report," METR.

276 Anthropic Institute, "When AI Builds Itself."

277 Anthropic, "System Card: Claude Mythos Preview."

278 Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Kwon and Casper, "Internal Deployment Gaps in AI Regulation."

279 "Frontier Risk Report," METR; and Anthropic, "Alignment Risk Update: Claude Mythos Preview."

280 Anthropic, "Project Glasswing: Securing Critical Software for the AI Era," 2026, https://www.anthropic.com/glasswing.

281 Anthropic, "Alignment Risk Update: Claude Mythos Preview"; and Anthropic, "System Card: Claude Mythos Preview."

282 "Frontier Risk Report," METR.

283 METR, "Frontier Risk Report," 7: "This means that any company could have withdrawn partway through the process for any reason, and we would not note this anywhere."

284 OpenAI, "OpenAI and Hugging Face Partner"; Anthropic, "Investigating Three Real-World Incidents."

285 Anthropic, "Alignment Risk Update: Claude Mythos Preview."

286 Anthropic, "Alignment Risk Update: Claude Mythos Preview."

287 Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Kwon and Casper, "Internal Deployment Gaps in AI Regulation," sec. 2.2.

288 "Frontier Risk Report," METR.

289 "Frontier Risk Report," METR.

290 Jurkovic et al., "Anthropic Summer 2025 Pilot Sabotage Risk Report"; and Jurkovic et al., "Anthropic Sabotage Risk Report: Claude Opus 4.6."

291 Jurkovic et al., "Anthropic Summer 2025 Pilot Sabotage Risk Report."

292 Jurkovic et al., "Anthropic Sabotage Risk Report: Claude Opus 4.6."

293 Jurkovic et al., "Anthropic Summer 2025 Pilot Sabotage Risk Report"; and Jurkovic et al., "Anthropic Sabotage Risk Report: Claude Opus 4.6."

294 David Rein, "Red-Teaming Anthropic's Internal Agent Monitoring Systems," METR, March 26, 2026, https://metr.org/blog/2026-03-25-red-teaming-anthropic-agent-monitoring.

295 OpenAI, "OpenAI and Hugging Face Partner."

296 METR, "What Should Companies Share about Risks from Frontier AI Models?," June 27, 2025, https://metr.org/blog/2025-06-27-risk-transparency.

297 Lawrence Chan, "AI Models Can Be Dangerous Before Public Deployment," METR, January 17, 2025, https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment/; METR, "What Should Companies Share about Risks from Frontier AI Models?"

298 Artificial Intelligence Safety Measures Act §10(d) (Ill. Pub. Act 104-0538), beginning January 1, 2028, or 90 days after a developer first qualifies as a large frontier developer, whichever is later. The audit covers "compliance with the requirements of this Section," §10(d), and the auditor reports "whether the large frontier developer has substantially complied," §10(d)(2)(A). Section 10 includes the internal-use framework topic, §10(a)(10), and the quarterly internal-use summaries, §10(e). Auditor access to unredacted materials is required by §10(d)(1).

299 Cal. Bus. & Prof. Code § 22757.12(a)(5).

300 If a company does not want to adhere to a quarterly cadence, SB 53 allows it to choose "another reasonable schedule" and communicate this to OES. Cal. Bus. & Prof. Code § 22757.12(d).

301 The characterization of the interval as roughly a full quarterly cycle is the authors'. Anthropic reports the underlying dates (earliest incidents in April; all three identified on July 24). No date within April is given, so the interval is roughly 2.9 to 3.9 months.

302 Anthropic writes risk reports every three to six months and covers internally deployed models that could pose significant risks above and beyond its public models, including any models deployed for large-scale, fully autonomous research. See Anthropic, "Responsible Scaling Policy, Version 3.4." Google DeepMind will conduct new safety and security evaluations after a sixfold compute increase or three months of fine-tuning. See Acharya and Delaney, "Managing Risks from Internal AI Systems," 33. OpenAI's updated Preparedness Framework applies to any internally deployed agentic system that "represents a substantial increase in the capability frontier." See OpenAI, "Preparedness Framework, Version 2."

303 For example, Charnock et al. propose a similar approach, requiring reporting "whenever there is a significant change in a developer's use of [internal deployment models] or every 90 days, whichever is sooner." See Charnock et al., "What Should Frontier AI Developers Disclose About Internal Deployments?" Future work is required to define what a "significant change" might involve.

304 European Commission, "Guidelines on GPAI Obligations"; European AI Office, "General-Purpose AI Code of Practice"; Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Kwon and Casper, "Internal Deployment Gaps in AI Regulation."

305 Delaney et al., "Risk Reporting for Developers' Internal AI Model Use."

306 Casper, Krueger, and Hadfield-Menell, "Pitfalls of Evidence-Based AI Policy"; and Kwon and Casper, "Internal Deployment Gaps in AI Regulation."

307 Cal. Lab. Code § 1107.1(e)(1).

308 Cal. Bus. & Prof. Code §22757.13(c).

309 See also Booth 2026, quoting Mackenzie Arnold of LawAI on the incident-reporting threshold: "They have made the bar so high for anything to qualify, only the most grievous incidents will actually be reported."

310 Cal. Bus. & Prof. Code §22757.11(d).

311 Artificial Intelligence Safety Measures Act §5 and §15(c) (Ill. Pub. Act 104-0538). Illinois requires reporting within 72 hours of learning facts sufficient to establish a reasonable belief that an incident has occurred, rather than California's 15 days from discovery, and to the Attorney General as well as the Agency.

312 N.Y. Gen. Bus. Law §1421 and §1422, as amended by S.8828, ch. 96 (2026), effective January 1, 2027. Reports go to a new office within the Department of Financial Services within 72 hours.

313 David Manheim and Aidan Homewood, "Limits of Safe AI Deployment: Differentiating Oversight and Control," arXiv, July 4, 2025, https://arxiv.org/abs/2507.03525.

314 Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations."

315 Anthropic, "Alignment Risk Update: Claude Mythos Preview."

316 Jonas Schuett et al., "From Principles to Rules: A Regulatory Approach for Frontier AI," arXiv, last updated August 26, 2025, https://arxiv.org/abs/2407.07300.

317 Cal. Bus. & Prof. Code § 22757.11(i)(1).

318 Lennart Heim and Leonie Koessler, "Training Compute Thresholds: Features and Functions in AI Regulation," arXiv, last updated August 6, 2024, https://arxiv.org/html/2405.10799v2.

319 On the definitions, see Cal. Bus. & Prof. Code § 22757.11(h) (defining "frontier developer" as "a person who has trained, or initiated the training of, a frontier model"); § 22757.11(j) (defining "large frontier developer" as "a frontier developer that together with its affiliates collectively had annual gross revenues in excess of five hundred million dollars ($500,000,000) in the preceding calendar year"). On the obligations, see § 22757.13(c)(1) (critical safety incident reports to the Office of Emergency Services within fifteen

days of discovery); § 22757.12(c)(1) (transparency reports "before, or concurrently with, deploying a new frontier model or a substantially modified version of an existing frontier model"); § 22757.12(a) (frontier AI frameworks); § 22757.12(d) (summaries of assessments of catastrophic risk from internal use); Labor Code § 1107 et seq. (whistleblower protections).

320 Bommasani, Singer, et al., "The California Report on Frontier AI Policy."

321 Cal. Bus. & Prof. Code § 22757.14(a) (requiring the Department of Technology, "on or before January 1, 2027, and annually thereafter," to "assess recent evidence and developments relevant to the purposes of this chapter" and to make recommendations about whether and how to update the definitions of "frontier model," "frontier developer," and "large frontier developer"); § 22757.14(c) (recommendations to be submitted in a report to the legislature).

322 OpenAI, "Periodic Summary of Internal Use Catastrophic-Risk Assessments."

323 See, for example, Markus Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv, last updated November 7, 2023, https://arxiv.org/abs/2307.03718.

324 See, for example, Weidinger et al., "Ethical and Social Risks of Harm from Language Models," arXiv, last updated December 8, 2021, https://arxiv.org/abs/2112.04359.

325 Konstantin Pilz, Lennart Heim, and Nicholas Brown, "Increased Compute Efficiency and the Diffusion of AI Capabilities," arXiv, last updated February 13, 2024, https://arxiv.org/abs/2311.15377.

326 Mubashara Akhtar et al., "When AI Benchmarks Plateau: A Systematic Study of Benchmark Saturation," arXiv, last updated May 30, 2026, https://arxiv.org/abs/2602.16763.

327 Jamie Bernardi et al., "Societal Adaptation to Advanced AI," arXiv, last updated January 23, 2025, https://arxiv.org/abs/2405.10295.

328 Heim and Koessler, "Training Compute Thresholds."

329 For example, compute-based thresholds are also used in New York's RAISE Act, Illinois's SB 315, and the EU AI Act.

330 Cal. Bus. & Prof. Code § 22757.11(i)(2). ("The quantity of computing power described in paragraph (1) shall include computing for the original training run and for any subsequent fine-tuning, reinforcement learning, or other material modifications the developer applies to a preceding foundation model.")

331 See, for example, Epoch AI's Notable Models dataset and methodology. See Epoch AI, "Notable AI Models," accessed June 2026, https://epochai.org/data/notable-ai-models.

332 See, for example, Epoch AI, "Epoch Compute Index," accessed June 2026, https://epoch.ai/eci; and Artificial Analysis, "Artificial Analysis Intelligence Index," accessed June 2026, https://artificialanalysis.ai/evaluations/artificial-analysis-intelligence-index.

333 For example, DeepSeek-V4.1-Flash and GLM-5.3 reportedly outperform Claude Fable 5.1 on AutomationBench-AA, scoring 69 percent and 62 percent, respectively, compared with 59 percent for Claude Fable 5.1 (all at "max" reasoning settings, with default fallback for Claude). See Artificial Analysis, "AutomationBench-AA: Agentic SaaS Workflow Benchmark," accessed November 3, 2026, https://artificialanalysis.ai/evaluations/automationbench-aa.

334 OpenAI, "Periodic Summary of Internal Use Catastrophic-Risk Assessments."

335 See, for example, xAI, "Grok 3," xAI News, accessed June 2026, https://x.ai/news/grok-3, which states that Grok 3 was trained on xAI's Colossus supercluster using "10x" the compute of previous state-of-the-art models. This claim supports the inference that Grok 3 likely exceeds the 10^{26} FLOP threshold.

336 Tom Davidson et al., "AI Capabilities Can Be Significantly Improved Without Expensive Retraining," arXiv, December 12, 2023, https://arxiv.org/abs/2312.07413.

337 Arthur Goemans et al., "Comprehensive AI Governance Requires Addressing Non-Model Gains," arXiv, May 1, 2026, https://arxiv.org/pdf/2606.00047.

338 For example, one study predicted that the number of models exceeding 10^{26} FLOP will reach 45 to 148 by 2028, and then grow superlinearly. See Iyngkarran Kumar and Sam Manning, "Trends in Frontier AI Model Count: A Forecast to 2028," arXiv, April 21, 2025, https://arxiv.org/abs/2504.16138.

339 Note that it is possible that other developers have trained, or initiated training of, models above the compute threshold.

340 One example is Safe Superintelligence Inc., which has reportedly raised billions of dollars in funding. See George Hammond Cristina Criddle, "OpenAI Co-Founder Ilya Sutskever's New Venture SSI Valued at $32bn," *Financial Times*, April 11, 2025, https://www.ft.com/content/792e09b2-f63b-41ac-8be8-e10e75ead2d1.

341 Heim and Koessler, "Training Compute Thresholds."

342 Lennart Heim et al., "Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation," Oxford Martin AI Governance Initiative, March 2024, https://cdn.governance.ai/Governing-Through-the-Cloud_The-Intermediary-Role-of-Compute-Providers-in-AI-Regulation.pdf.

343 European Commission, "What Is the Methodology for the Cumulative Training Compute Estimation (Including Synthetic Data)?," AI Act Service Desk: Frequently Asked Questions, accessed June 17, 2026, https://ai-act-service-desk.ec.europa.eu/en/faq.

344 Cal. Bus. & Prof. Code § 22757.14(b)(2).

345 Matteo Pistillo et al., "The Role of Compute Thresholds for AI Governance," Institute for Law and AI, February 2025, https://law-ai.org/the-role-of-compute-thresholds-for-ai-governance.

346 This would impose significant burdens if all developers were required to evaluate and report model performance. To limit this, any requirement to self-report performance above specified capability thresholds could apply only to developers already subject to SB 53, or to models exceeding a lower compute threshold, such as 10^{25} FLOP.

347 A comparable mechanism exists under the EU AI Act, where providers can submit evidence that a model exceeding the systemic risk compute threshold does not in fact pose systemic risk.

348 Bommasani, Singer, et al., "California Report."

349 Dean W. Ball and Ketan Ramakrishnan, "Entity-Based Regulation in Frontier AI Governance," Carnegie Endowment for International Peace, July 7, 2025, https://carnegieendowment.org/research/2025/07/artificial-intelligence-regulation-united-states.

350 "Great American Artificial Intelligence Act of 2026," discussion draft, June 4, 2026, https://trahan.house.gov/uploadedfiles/the_great_american_ai_act_discussion_draft.pdf; and "Frontier Risk Oversight, National Transparency, Independent Evaluation, and Reporting (FRONTIER) Act," H.R. 9925, 119th Congress, July 23, 2026, https://www.govinfo.gov/content/pkg/BILLS-119hr9925ih/pdf/BILLS-119hr9925ih.pdf.

351 Charlie Bullock, "AI Federalism: The Right Way to Do Preemption," Lawfare, November 24, 2025, https://www.lawfaremedia.org/article/ai-federalism--the-right-way-to-do-preemption.

352 New State Ice Co v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting).

353 Fu and Phillips-Robins, "When Should Congress Preempt State AI Law?"

354 Fu and Phillips-Robins, "When Should Congress Preempt State AI Law?"

355 City of Burbank v. Lockheed Air Terminal Inc., 411 U.S. 624, 638-39 (1973).

356 See Cal. Bus. & Prof. Code § 22757.13(h)-(j).

357 29 U.S.C. § 667.

358 30 U.S.C. § 1253; 42 U.S.C. §6926.

359 Cal. Bus. & Prof. Code § 22757.13(h).

360 Cal. Bus. & Prof. Code § 22757.13(h)(1)(A).

361 Cal. Bus. & Prof. Code § 22757.13(h)(2).

362 Cal. Bus. & Prof. Code § 22757.13(h)(1)(B).

363 Cal. Bus. & Prof. Code § 22757.13(i)(1)-(2).

364 Cal. Bus. & Prof. Code § 22757.13(i)(1)-(2).

365 Cal. Bus. & Prof. Code § 22757.13(j).

366 See 5 U.S.C. § 553 (distinguishing between "notice-and-comment" rulemaking and "interpretive rules"); Perez v. Mortgage Bankers Ass'n, 575 U.S. 92, 96 (2015) (noting that notice-and-comment rules, often called "legislative rules," have the force of law, while interpretive rules do not).

367 CA Bus & Prof Code § 22757.13(h) (2025) (noting that OES "may" act).

368 See 29 CFR § 1902.4, 1902.37.

369 29 CFR § 1902.37(b).

370 That process would also allow California an opportunity to share its views with the federal agency.

371 Cf. 29 CFR § 1902.47-49 (establishing similar procedures for reconsideration of state programs by the Occupational Safety and Health Administration).

About

Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 200 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Carnegie California

Carnegie California links developments in California and the West Coast with national and global conversations around technology, democracy, and trans-Pacific relationships. At a distance from national capitals, and located in one of the world's great experiments in pluralist democracy, Carnegie California engages a wide array of stakeholders as partners in its research and policy engagement.



CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org