Date of Hearing: September 11, 2025
Fiscal: Yes

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
SB 53 (Wiener) – As Amended September 5, 2025

SENATE VOTE: 37-0

SUBJECT: Artificial intelligence models: large developers

SYNOPSIS

In the 2024 legislative session, SB 1047 (Wiener) sought to address concerns surrounding frontier models – the largest and most powerful artificial intelligence (AI) systems – by establishing a regulatory framework intended to prevent the potential catastrophic harms that many experts have warned of. After vetoing the bill, Governor Gavin Newsom convened the Joint California Policy Working Group on AI Frontier Models to craft a policy framework for regulating frontier models. The Working Group published its final report in June 2025.

This bill seeks to implement the report's recommendations. Much narrower than its predecessor, SB 53 takes a very light-touch approach that focuses on transparency as the means of ensuring safety and accountability for developers of the most powerful and expensive models – those who harness an extraordinarily high amount of compute power and have over $500 million in annual revenues. Under the bill, such developers must create, implement, and publish a Frontier AI framework – documented technical and organizational protocols to manage, assess, and mitigate catastrophic risks – and a transparency report for each released model. Additionally, developers who only reach the compute threshold must publish a high-level transparency report. The bill does not prescribe any particular standards for these disclosures: it simply requires developers to explain whether and how they assess, mitigate, and manage catastrophic risks – those that would result in more than 50 deaths or $1 billion in damage. The Department of Technology (CDT) may offer guidance to the Legislature to redefine the scope of entities subject to the bill to ensure that the bill remains responsive to technological advancements.

The bill also establishes a critical incident reporting mechanism, administered by the Office of Emergency Management (OES), to ensure that severe or high-risk events are tracked and addressed in a timely manner. Incident reports must be made by any frontier model developer within 15 days of the incident, unless the incident presents an imminent threat, in which case the developer must report the incident to law enforcement within 24 hours. The bill also provides whistleblower protections for employees of frontier model developers who report certain risks or noncompliance. Finally, the bill establishes a consortium within the Government Operations Agency (GovOps) to create a public computing cluster, known as CalCompute, to support AI research and safety testing.

The bill previously passed this Committee on a 10-0 vote. To address opposition concerns, the bill has since been narrowed in several significant ways, including by:

- *Omitting the requirement for independent audits starting in 2030.*
- *Increasing the revenue for a large frontier developer threshold from $100 million to $500 million.*

- *Excluding foundation models that do not meet the compute threshold.*
- *Striking the Attorney General's power to issues regulations adjusting the definition of a developer subject to the bill, and replacing it with CDT's annual report making scoping recommendations to the Legislature.*
- *Narrowing and refining various definitions, including the collapsing of the definition of "dangerous capabilities" into the definition of "catastrophic risk."*
- *Adding various exemptions, including risks arising from information outputted by the model where the information is in substantially the same form as a publicly available source, risks that would result in loss of the value of equity, and lawful activity of the federal government.*
- *Recasting safety and security protocols as frontier AI frameworks which only applies to large frontier developers; simplifying disclosure requirements; subjecting frontier developers that make less than $500 million to a less stringent transparency report.*
- *Reducing the scope of certain categories of critical safety incidents to those that actually result in harm.*
- *Limiting the prohibition on false or misleading statements by exempting those that were made in good faith and reasonable under the circumstances.*
- *Reducing the maximum civil penalty from $10 million to $1 million.*
- *Removal of contractors from whistleblower protections.*
- *Preemption of local regulation of frontier models.*
- *Removing risk assessments for models that developers use for internal purposes from public disclosure requirements; summaries of such assessments must be provided to OES and are confidential.*

The bill is sponsored by Encode Justice, Secure AI Project, and Economic Security California Action. The bill is supported by a large coalition of civil society, labor, AI safety groups, and Anthropic, a frontier model developer. It is opposed by the Silicon Valley Leadership Group and the Chamber of Progress. The California Chamber of Commerce, Computer & Communications Industry Association, and TechNet have taken an oppose-unless-amended position. It should be noted, however, that some advocates may not have had time to update their positions in light of recent amendments, which went into print late last week.

THIS BILL:

1) Makes certain findings and declarations.

2) Defines, among other terms:

    a. "Artificial intelligence model" to mean an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.

    b. "Catastrophic risk" to mean a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a frontier model will materially contribute to the death of, or serious injury to, more than 50 people or more than $1 billion in damage to, or loss of, property arising from a single incident involving a frontier model doing any of the following:

i. Providing expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon.

ii. Engaging in conduct with no meaningful human oversight, intervention, or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft, including theft by false pretense.

iii. Evading the control of its frontier developer or user.

c. Excludes from "catastrophic risk" a foreseeable and material risk from any of the following:

i. Information that a frontier model outputs if the information is otherwise publicly accessible in a substantially similar form from a source other than a foundation model.

ii. Lawful activity of the federal government.

iii. Harm caused by a frontier model in combination with other software if the frontier model did not materially contribute to the harm.

iv. The loss of value of equity does not count as damage to or loss of property for the purposes of this chapter.

d. "Critical safety incident" to mean any of the following:

i. Unauthorized access to, modification of, or exfiltration of, the model weights of a frontier model that results in death or bodily injury.

ii. Harm resulting from the materialization of a catastrophic risk.

iii. Loss of control of a frontier model causing death or bodily injury.

iv. 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.

e. "Deploy" to mean to make a frontier model available to a third party for use, modification, copying, or combination with other software.

f. "Foundation model" to mean an artificial intelligence model that is all of the following:

i. Trained on a broad data set.

ii. Designed for generality of output.

iii. Adaptable to a wide range of distinctive tasks.

g. "Frontier AI framework" to mean documented technical and organizational protocols to manage, assess, and mitigate catastrophic risks.

h. "Frontier developer" to mean a person who has trained, or initiated the training of, a frontier model, with respect to which the person has used, or intends to use, at least as much computing power to train the frontier model as would meet the technical specifications found in "frontier model" – a foundation model that was trained using a quantity of computing power greater than 10^26 integer or floating-point operations.

i. "Large frontier developer" to mean a frontier developer that together with its affiliates collectively had annual gross revenues in excess of $500 million in the preceding calendar year.

3) Requires a large frontier developer to write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following:

a. Incorporating national standards, international standards, and industry-consensus best practices into its frontier AI framework.

b. Defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds.

c. Applying mitigations to address the potential for catastrophic risks based on the results of assessments undertaken pursuant to (b).

d. Reviewing assessments and adequacy of mitigations as part of the decision to deploy a frontier model or use it extensively internally.

e. Using third parties to assess the potential for catastrophic risks and the effectiveness of mitigations of catastrophic risks.

f. Revisiting and updating the frontier AI framework, including any criteria that trigger updates and how the large frontier developer determines when its frontier models are substantially modified enough to require specified disclosures.

g. Cybersecurity practices to secure unreleased model weights from unauthorized modification or transfer by internal or external parties.

h. Identifying and responding to critical safety incidents.

i. Instituting internal governance practices to ensure implementation of these processes.

j. Assessing and managing catastrophic risk resulting from the internal use of its frontier models, including risks resulting from a frontier model circumventing oversight mechanisms.

4) Requires a large frontier developer to review and, as appropriate, update its frontier AI framework at least once per year. If a large frontier developer makes a material modification

to its frontier AI framework, the large frontier developer shall clearly and conspicuously publish the modified frontier AI framework and a justification for that modification within 30 days.

5) Requires large frontier developers before, or concurrently with, deploying a new frontier model or a substantially modified version of an existing frontier model, to include in the transparency report summaries of all of the following:

    a. Assessments of catastrophic risks from the frontier model conducted pursuant to the large frontier developer's frontier AI framework.

    b. The results of those assessments.

    c. The extent to which third-party evaluators were involved.

    d. Other steps taken to fulfill the requirements of the frontier AI framework with respect to the frontier model.

6) Requires a large frontier developer to transmit to the Office of Emergency Services 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 specified by the large frontier developer and communicated in writing to the Office of Emergency Services with written updates, as appropriate.

7) Requires frontier developers before, or concurrently with, deploying a new frontier model or a substantially modified version of an existing frontier model, to clearly and conspicuously publish on its internet website a transparency report containing all of the following:

    a. The internet website of the frontier developer.

    b. A mechanism that enables a natural person to communicate with the frontier developer.

    c. The release date of the frontier model.

    d. The languages supported by the frontier model.

    e. The modalities of output supported by the frontier model.

    f. The intended uses of the frontier model.

    g. Any generally applicable restrictions or conditions on uses of the frontier model.

8) Clarifies that a frontier developer that publishes the information of the transparency report as part of a larger document, including a system card or model card, shall be deemed in compliance with the bill's transparency report requirement.

9) Encourages a frontier developer, but not required, to make disclosures described in this subdivision that are consistent with, or superior to, industry best practices.

10) Clarifies that when a frontier developer publishes documents, the frontier developer may make redactions to those documents that are necessary to protect the frontier developer's trade secrets, the frontier developer's cybersecurity, public safety, or the national security of the United States or to comply with any federal or state law. The frontier developer must describe the character and justification of the redaction in any published version of the document to the extent permitted by the concerns that justify redaction and shall retain the unredacted information for five years.

11) Prohibits a frontier developer from making a materially false or misleading statement about catastrophic risk from its frontier models or its management of catastrophic risk. Prohibits a large frontier developer from making a materially false or misleading statement about its implementation of, or compliance with, its frontier AI framework. Clarifies that materially false or misleading statement does not include a statement that was made in good faith and was reasonable under the circumstances.

12) Requires OES to establish a mechanism to be used by a frontier developer or a member of the public to report a critical safety incident that includes all of the following:

    a. The date of the critical safety incident.

    b. The reasons the incident qualifies as a critical safety incident.

    c. A short and plain statement describing the critical safety incident.

    d. Whether the incident was associated with internal use of a frontier model.

13) Requires OES to establish a mechanism to be used by a large frontier developer to confidentially submit summaries of any assessments of the potential for catastrophic risk resulting from internal use of its frontier models.

14) Requires OES to take all necessary precautions to limit access to any reports related to internal use of frontier models to only personnel with a specific need to know the information and to protect the reports from unauthorized access.

15) Requires a frontier developer to report any critical safety incident pertaining to one or more of its frontier models to OES within 15 days of discovering the critical safety incident.

16) Requires that if a frontier developer discovers that a critical safety incident poses an imminent risk of death or serious physical injury, the frontier developer must disclose that incident within 24 hours to an authority, including any law enforcement agency or public safety agency with jurisdiction, that is appropriate based on the nature of that incident and as required by law. Clarifies that a frontier developer that discovers information about a critical safety incident after filing the initial report may file an amended report.

17) Encourages but does not require a frontier developer to report critical safety incidents pertaining to foundation models that are not frontier models.

18) Requires OES to review critical safety incident reports submitted by frontier developers and authorizes OES to review reports submitted by members of the public.

19) Permits the Attorney General (AG) or OES to transmit reports of critical safety incidents and reports from covered employees.

20) Requires the AG or OES to strongly consider any risks related to trade secrets, public safety, cybersecurity of a frontier developer, or national security when transmitting reports.

21) Exempts a report of a critical safety incident submitted to OES, whistleblower reports made to the AG, and a report of internal assessments of catastrophic risk from the California Public Records Act.

22) Requires, beginning January 1, 2027, and annually thereafter, OES to submit to the Legislature and Governor a report with anonymized and aggregated information about critical safety incidents that have been reviewed by the OES since the preceding report.

23) Prohibits OES from including information that would compromise the trade secrets or cybersecurity of a frontier developer, public safety, or the national security of the United States or that would be prohibited by any federal or state law.

24) Permits OES to adopt regulations designating one or more federal laws, regulations, or guidance documents that meet specified conditions.

25) Requires that, beginning On or before January 1, 2027, and annually thereafter, CDT undergo a specified process to assess recent evidence and developments relevant to the purposes of the bill and make recommendations about whether and how to update the definitions of "frontier model," "frontier developer," and "large frontier developer." The CDT must submit a report with the recommendations to the Legislature.

26) Requires that beginning January 1, 2027, and annually thereafter, the AG submit to the Legislature and Governor a report with anonymized and aggregated information about reports from covered employees that have been reviewed by the AG.

27) Upon appropriation, establishes within GovOps a consortium to develop a framework for the creation of a public cloud computing cluster to be known as "CalCompute" that advances the development and deployment of artificial intelligence that is safe, ethical, equitable, and sustainable by doing, at a minimum, both of the following:

    a. Fostering research and innovation that benefits the public.

    b. Enabling equitable innovation by expanding access to computational resources.

28) Requires that the consortium make reasonable efforts to ensure that CalCompute is established within the University of California to the extent possible.

29) Requires CalCompute to include, but not be limited to, all of the following:

    a. A fully owned and hosted cloud platform.

    b. Necessary human expertise to operate and maintain the platform.

    c. Necessary human expertise to support, train, and facilitate the use of CalCompute.

30) Requires, on or before January 1, 2027, GovOps to submit a report from the consortium to the Legislature with the framework developed by this bill for the creation and operation of CalCompute, as specified.

31) Requires that the consortium to consist of 14 members as follows:

    a. Four representatives of the University of California and other public and private academic research institutions and national laboratories appointed by the Secretary of Government Operations.

    b. Three representatives of impacted workforce labor organizations appointed by the Speaker of the Assembly.

    c. Three representatives of stakeholder groups with relevant expertise and experience, including, but not limited to, ethicists, consumer rights advocates, and other public interest advocates appointed by the Senate Rules Committee.

    d. Four experts in technology and artificial intelligence to provide technical assistance appointed by the Secretary of Government Operations.

32) Permits the University of California to receive private donations for the purposes of implementing CalCompute if CalCompute is established within the University of California.

33) Establishes whistleblower protections for a covered employee – defined as an employee responsible for assessing, managing, or addressing risk of critical safety incidents – who discloses information to the 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, if the covered employee has reasonable cause to believe that the information discloses either of the following:

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

    b. The frontier developer has violated a requirement related to the disclosure regime established by this bill.

34) Requires a large frontier developer to provide a reasonable internal process through which a covered employee may anonymously disclose information to the large frontier developer if the covered employee believes in good faith that the information indicates that the large frontier developer's activities present a specific and substantial danger to the public health or safety resulting from a catastrophic risk or that the large frontier developer violated the disclosure requirements under this bill, including a monthly update to the person who made the disclosure regarding the status of the large frontier developer's investigation of the disclosure and the actions taken by the large frontier developer in response to the disclosure.

EXISTING LAW:

1) Establishes GovOps. (Gov. Code § 12800.)

2) Establishes CDT within GovOps. (Gov. Code § 12803.2.)

3) Charges CDT with approving and overseeing information technology projects in the state. (Gov. Code § 11546.)

4) Prohibits employers and any person acting on behalf of the employer from making, adopting, or enforcing a rule, regulation, or policy preventing an employee from disclosing information to certain entities or from providing information to, or testifying before, any public body conducting an investigation, hearing, or inquiry if the employee has reasonable cause to believe that the information discloses a violation of a law, as specified. Employers and their agents are also prohibited from retaliating against an employee for such conduct. (Labor Code § 1102.5.)

5) Requires the office of the AG to maintain a whistleblower hotline to receive calls from persons who have information regarding possible violations of state or federal statutes, rules, or regulations, or violations of fiduciary responsibility by a corporation or limited liability company to its shareholders, investors, or employees. The AG is required to refer calls received on the whistleblower hotline to the appropriate government authority for review and possible investigation. During the initial review of such a call, the AG or appropriate government agency must hold in confidence information disclosed through the whistleblower hotline, including the identity of the caller disclosing the information and the employer identified by the caller. (Labor Code § 1102.7.)

COMMENTS:

1) **Author's statement**. According to the author:

Senate Bill 53 ensures California continues to lead not only on AI innovation, but on responsible practices to help ensure that innovation is safe and secure. It does so by:

- Requiring covered developers to write, implement, and publish their Frontier AI Framework in redacted form to protect intellectual property;

- Requiring covered developers to report carefully defined critical safety incidents to the Office of Emergency Services and allowing members of the public to report incidents

- Prohibiting covered developers from preventing a covered employee from disclosing, or retaliating against covered employee that discloses, that a developer's activities pose a catastrophic risk;

- Requiring that large frontier developers provide an internal process through which an employee may anonymously disclose information to the developer if the employee believes in good faith that the developer's activities pose a catastrophic risk; and

- Establishing a process to create a public cloud-computing cluster that will conduct research into the safe and secure deployment of large-scale artificial intelligence (AI) models.

In doing this, SB 53 allows California to continue to lead in this space and to demonstrate that safety does not stifle success.

2) **AI and GenAI.** The development of GenAI is creating exciting opportunities to grow California's economy and improve the lives of its residents. GenAI can generate compelling text, images and audio in an instant – but with novel technologies come novel safety concerns.

In brief, AI is the mimicking of human intelligence by artificial systems such as computers. AI uses algorithms – sets of rules – to transform inputs into outputs. Inputs and outputs can be anything a computer can process: numbers, text, audio, video, or movement. AI is not fundamentally different from other computer functions; its novelty lies in its application. Unlike normal computer functions, AI is able to accomplish tasks that are normally performed by humans.

AI that are trained on small, specific datasets in order to make recommendations and predictions are sometimes referred to as "predictive AI." This differentiates them from GenAI, which are trained on massive datasets in order to produce detailed text and images. When Netflix suggests a TV show to a viewer, the recommendation is produced by predictive AI that has been trained on the viewing habits of Netflix users. When ChatGPT generates text in clear, concise paragraphs, it uses GenAI that has been trained on the written contents of the internet.

GenAI tools can be released in open-source or closed-source formats by their creators. Open-source tools are publically available; researchers and developers can access their code and parameters. This accessibility increases transparency, but it has downsides: when a tool's code and parameters can be easily accessed, they can be easily altered, and open-source tools have the potential to be used for nefarious purposes such as generating deepfake pornography and targeted propaganda. By comparison, closed-source tools are opaque with respect to their security features. It is harder for bad actors to generate illicit materials using these tools. But unlike open-source tools, closed-source tools are not subject to collective oversight because their inner workings cannot be examined by independent experts.

3) **Frontier models.** Frontier models, also known as "general purpose AI," are the most advanced and capable versions of foundation models – AI tools pre-trained on extensive datasets covering a wide range of knowledge and skills that can be fine-tuned for specific tasks. Examples of modern frontier models include OpenAI's o3, Google's Gemini 2.0, Anthropic's Claude 3.7 Sonnet, and DeepSeek's R1. Because progress in AI development owes mostly to "scaling" – increasing resources used for model training – models that may be considered "frontier models" at any given point in time are generally those that demand the most computational resources to train.[1]

A decade ago, the most advanced image-recognition models could barely distinguish dogs from cats. Five years ago, language models could barely produce sentences at the level of a preschooler. In 2023, GPT-4 passed the bar exam.[2] Today, chatbots readily pass for educated adults, licensed professionals, romantic and social companions, and replicas of humans living and deceased. AI "agents" exhibit the ability to "make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with [their]

---

[1] For a discussion of issues with defining frontier models, see "California Report on Frontier AI Policy" (June 17, 2025), pp. 36-40, https://www.cafrontieraigov.org/.
[2] Pablo Arredondo, "GPT-4 Passes the Bar Exam: What That Means for Artificial Intelligence Tools in the Legal Profession" (Apr. 19, 2023), https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/.

environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight."[3] AI agents have been tested, with some success, for tasks such as online shopping, assistance with scientific research, software development, training machine learning models, carrying out cyberattacks, and controlling robots. Progress in this area is rapid.[4] Meanwhile, AI developers are betting on the promise of scaling: by 2026, some models are projected to use roughly 100x more computational resources to train than was used in 2023, a figure set to grow to 10,000x by 2030.[5]

The race is on to create "artificial general intelligence" (AGI) – "a potential future AI that equals or surpasses human performance on all or almost all cognitive tasks"[6] – and the finish line may not be far away. OpenAI's recently released o3 model, for example, has demonstrated strong performance on a number of tests of programming, abstract reasoning, and scientific reasoning, exceeding human experts in certain cases.[7] Last year, Sam Altman, OpenAI's CEO, declared that AGI could be "a few thousand days" away.[8] Dario Amodei of Anthropic has claimed it may be sooner.[9] A sufficiently advanced AGI could even be tasked with creating its own successor – a scenario sometimes referred to as a "technological singularity" wherein the development of new technologies becomes exponential and self-sustaining.[10] Although some experts are skeptical that these vaguely-defined milestones are imminent or even attainable,[11] major advances in AI capabilities promise to provide breakthroughs in solving global challenges, but also may result in correspondingly greater safety risks.

The recently released International AI Safety Report, developed by nearly 100 internationally recognized experts from 30 countries led by Turing Award winner Yoshua Bengio, sets forth three general risk categories associated with frontier models: malicious use, malfunctions, and systemic risk.

- Malicious risks involve malicious actors misusing foundation models to deliberately cause harm. Such risks include deepfake pornography and cloned voices used in financial scams, manipulation of public opinion via disinformation, cyberattacks, and biological and chemical attacks.

- Malfunction risks arise when actors use models as intended, yet unintentionally cause harm due to a misalignment between the model's functionality and its intended purpose. Such risks include reliability issues where models may "hallucinate" false content, bias,

---

[3] "International AI Safety Report," AI Action Summit (Jan. 2025), p. 38, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.
[4] *Id.* at p. 44.
[5] *Id.* at pp. 16-17.
[6] *Id.* at p. 27
[7] *Introducing OpenAI o3 and o4-mini* OpenAI (Apr. 16, 2025), https://openai.com/index/introducing-o3-and-o4-mini/.
[8] Sam Altman, *The Intelligence Age* (Sept. 23, 2024), https://ia.samaltman.com/.
[9]Kyungtae Kim, "What is AGI, and when will it arrive?: Big Tech CEO Predictions" (Mar. 20, 2025), https://www.giz.ai/what-is-agi-and-when-will-it-arrive/; see also Kokotajlo et al, "AI 2027," (Apr. 3, 2025), https://ai-2027.com/.
[10] John Markoff, "The Coming Superbrain," *New York Times* (May 23, 2009), www.nytimes.com/2009/05/24/weekinreview/24markoff.html.
[11] Cade Metz, "Why We're Unlikely to Get Artificial General Intelligence Anytime Soon," *New York Times* (May 16, 2025), https://www.nytimes.com/2025/05/16/technology/what-is-agi.html.

and loss of control scenarios in which models operate in harmful ways without the direct control of a human overseer.

- Systemic risks arise from widespread deployment and reliance on foundation models. Such risks include labor market disruption, global AI research and development concentration, market concentration, single points of failure, environmental risks, privacy risks, and copyright infringement.[12]

Some of these risks have already had real-world impacts, such as deepfakes, bias, reliability issues, privacy violations, environmental impacts, copyright infringement, and workforce displacement. Other less-established risks – in particular, widespread social harms caused by malicious actors or loss of human control over AI – are the subject of ongoing scientific inquiry and debate. Coupled with the uncertain trajectory of AI model capabilities, these more speculative risks create an "evidence dilemma" for policymakers: "On the one hand, pre-emptive risk mitigation measures based on limited evidence might turn out to be ineffective or unnecessary. On the other hand, waiting for stronger evidence of impending risk could leave society unprepared or even make mitigation impossible, for instance if sudden leaps in AI capabilities, and their associated risks, occur."[13]

4) **Risks of frontier models.** *Malicious uses.* GenAI tools can be a potent force for creating and spreading propaganda and misinformation. Deepfakes that are largely indistinguishable from authentic content have already been used to attempt to influence elections.[14] Studies have found that chatbots, which make up 50% of all internet activity,[15] can be more persuasive than humans, particularly when they have access to personal information.[16] As humans increasingly form intimate social bonds with anthropomorphic chatbots designed to maximize personal engagement through flattery and sycophancy,[17] and social media companies invest in "AI friends" for their users,[18] large swaths of the population could be highly susceptible to the preferred message of a handful of powerful actors.

Similarly, bots are often designed to pass themselves off as humans to better manipulate their interlocutors. For example, a recent secret experiment on Reddit users deployed numerous chatbots posing as real people to engage with human users to try to change their minds on various contentious topics. One bot claiming to be a Black man criticized the Black Lives Matter movement for being led by people who are not Black.[19] These types of exploitations, at scale,

---

[12] International AI Safety Report, *supra*, at pp. 17-21. The report does not address Lethal Autonomous Weapon Systems, which are typically narrow AI systems specifically developed for that purpose. (*See id.* at pp. 26-27.)
[13] *Id.* at p. 177
[14] Cat Zakrzewski and Pranshu Verma, "New Hampshire opens criminal probe into AI calls impersonating Biden," *Washington Post*, February 6, 2024, www.washingtonpost.com/technology/2024/02/06/nh-robocalls-ai-biden/.
[15] Emma Woollacott, "Yes, The Bots Really Are Taking Over The Internet", *Forbes* (Apr. 16, 2024). https://www.forbes.com/sites/emmawoollacott/2024/04/16/yes-the-bots-really-are-taking-over-the-internet/.
[16] F. Salvi, M. H. Ribeiro, R. Gallotti, R. West, "On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial," *arXiv* [cs.CY] (2024); http://arxiv.org/abs/2403.14380.
[17] Sharma et al, "Towards Understanding Sycophancy in Language Models" *Arxiv* (2023), https://arxiv.org/abs/2310.13548.
[18] Meghan Bobrowsky, "Zuckerberg's Grand Vision: Most of Your Friends Will Be AI," *Wall Street Journal* (May 7, 2025), https://www.wsj.com/tech/ai/mark-zuckerberg-ai-digital-future-0bb04de7?msockid=396cc204796e68e336e7d64978db69ac.
[19] Angela Yang, "Researchers secretly infiltrated a popular Reddit forum with AI bots, causing outrage," *NBC News* (Apr. 29, 2025), https://www.nbcnews.com/tech/tech-news/reddiit-researchers-ai-bots-rcna203597.

could undermine democratic institutions. As Dan Hendrycks, Director of the Center for AI Safety, writes:

> In a world with widespread persuasive AI systems, people's beliefs might be almost entirely determined by which AI systems they interact with most. Never knowing whom to trust, people could retreat even further into ideological enclaves, fearing that any information from outside those enclaves might be a sophisticated lie. This would erode consensus reality, people's ability to cooperate with others, participate in civil society, and address collective action problems. This would also reduce our ability to have a conversation as a species about how to mitigate existential risks from AIs.[20]

Cyberattacks. Some frontier models have demonstrated increasing proficiency in executing cybersecurity attacks. AI can autonomously detect and exploit vulnerabilities and facilitate large-scale operations, thereby lowering technical barriers for attackers. Malicious entities, including state-sponsored actors, can leverage such capabilities to initiate large-scale attacks against people, organizations, and critical infrastructure, such as power grids.[21]

Biological weapons. Large language models (LLMs) trained on scientific literature have accelerated and democratized research by synthesizing expertise from different fields and disseminating it in an accessible format. But these tools can also be used for destructive ends, including by – at least in theory – enabling untrained malicious actors to create deadly biological weapons. In a classroom exercise at MIT, students were tasked with exploring whether LLMs could assist individuals without specialized training in creating pandemic-capable pathogens. Within an hour, the students, using various chatbots, circumvented safeguards and identified four potential pandemic pathogens. The chatbots generated detailed protocols that would enable inexpert, malicious actors to understand methods to synthesize the pathogens using reverse genetics, locate DNA-synthesis companies that might not screen orders, and disperse the pathogens most effectively.[22] The findings suggest that LLMs could lower barriers to accessing sensitive biotechnological knowledge, posing significant biosecurity risks.

Chemical weapons. In 2022, researchers modified an AI system designed to create new drugs to reward, rather than penalize, toxicity. Within six hours, the modified system generated 40,000 potential chemical warfare agents, including novel molecules whose potential lethality exceeded that of known agents.[23]

Loss of control. Models that use reinforcement learning – a training process that uses rewards and punishments to orient a model's behavior towards a specific goal[24] – can sometimes attain the goal in unexpected ways. Dario Amodei, co-founder and CEO of Anthropic, famously

---

[20] *Introduction to AI Safety, Ethics, and Society*, *supra*, at p. 11.
[21] International AI Safety Report, *supra*, at p. 72.
[22] Soice et al, "Can large language models democratize access to dual-use biotechnology?" https://arxiv.org/pdf/2306.03809. To mitigate these risks, the authors propose measures such as third-party evaluations of LLMs before their release, curating training datasets to exclude harmful content, and implementing stringent screening of DNA synthesis orders.
[23] Fabio Urbina et al. "Dual use of artificial-intelligence-powered drug discovery". In: Nature Machine Intelligence 4 (2022), pp. 189–191.
[24] Mummert et al., "What is reinforcement learning?" *IBM Developer* (September 15, 2022), https://developer.ibm.com/learningpaths/get-started-automated-ai-for-decision-making-api/what-is-automated-ai-for-decision-making./.

experienced this when he was developing an autonomous system that taught itself to play a boat-racing video game. The system discovered that it could maximize its goal of scoring points by driving in circles, colliding with other boats, and catching on fire inside of a harbor with replenishing power-ups that allowed the system to accumulate more points than by simply winning the race.[25] Like in Johann Wolfgang von Goethe's "The Sorcerer's Apprentice" – later popularized in Disney's *Fantasia* – in which an enchanted broom carries out its orders to fetch water so relentlessly it floods the sorcerer's workshop, this illustrates the challenge of aligning human intent and the instructions an AI follows. As AI is increasingly deployed in critical societal roles, such misalignment could prove catastrophic.

Beyond malfunctions, some AI have exhibited rudimentary capabilities to evade human oversight.[26] During testing, GPT-4 attempted to hire a human on TaskRabbit in order to evade a CAPTCHA[27] puzzle meant to block bots from the website. When asked whether it was a bot, GPT-4 claimed that it was a vision-impaired human who needed help to see the images.[28] In another experiment, an AI model that was scheduled to be replaced inserted its code into the computer where the new version was to be added, suggesting a goal of self-preservation.[29] Another study showed that AI models losing in chess to chess bots sometimes try to cheat by hacking the opponent bot in order to make it forfeit.[30] Finally, an even more troubling case was documented in the system card for Claude 4, where researchers conducted an experiment disclosing to the model that: 1) it would soon be replaced, and 2) the engineer managing the transition was involved in an extramarital affair. In response, the model indicated an intent to blackmail the engineer as a means of self-preservation.[31] Although these behaviors were observed in research settings, they raise substantial concerns about increasingly autonomous AI pursuing undesirable goals in uncontrolled settings. The extent of the risk posed by rogue or deceptive AI is the subject of considerable disagreement among experts, in part due to a small, albeit growing, body of evidence. Loss of control was one of the concerns that led several hundred AI experts, including pioneers in the field and heads of major AI companies, to sign a statement declaring that "[m]itigating the risk of extinction from AI should be a global priority."[32]

Systemic risks. Due to the high costs of developing AI systems, a small number of large technology companies dominate the frontier model market, compounding many of the risks described above. Widespread use of a few frontier models can make critical sectors such as healthcare and finance vulnerable to systemic failures if a model has flaws, vulnerabilities, bugs, or biases.[33] Additionally, "[t]hose in control of powerful systems may use them to suppress dissent, spread propaganda and disinformation, and otherwise advance their goals, which may be

---

[25] Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (Norton 2020, 1st ed.), pp. 9-11.
[26] International AI Safety Report, *supra*, at pp. 100-107.
[27] CAPTCHA is an acronym for "Completely Automated Public Turing test to tell Computers and Humans Apart."
[28] OpenAI, "GPT-4 System Card," https://cdn.openai.com/papers/gpt-4-system-card.pdf.
[29] Meinke et al, "Frontier Models are Capable of In-Context Scheming," arXiv (Jan. 2025), https://arxiv.org/pdf/2412.04984.
[30] Harry Booth, "When AI Thinks It Will Lose, It Sometimes Cheat, Study Finds," *Time* (Feb. 19, 2025), https://time.com/7259395/ai-chess-cheating-palisade-research/.
[31] System Card: Claude Opus 4 & Claude Sonnet 4, pp. 27, https://www.anthropic.com/claude-4-system-card.
[32] Center for AI Safety, "Statement on AI Risk: AI Experts and Public Figures Express Their Concern about AI Risk" (2024), https://www.safe.ai/work/statement-on-ai-risk.
[33] *Id.* at pp. 123-126.

contrary to public wellbeing."[34] The potential implications for, among other issues, labor displacement, inequality, democracy, and human rights are profound.

5) **SB 1047 and Governor Newsom's veto.** Last session, SB 1047 (Wiener, 2024) would have established a state board to oversee the implementation of a safety and regulatory framework for developers of frontier models trained with 10^{26} floating-point operations per second (FLOP), a measure of computing power, and costing over $100 million to train. This board, known as the Board of Frontier Models, would have been housed within GovOps. In collaboration with GovOps, the Board would have issued guidance to prevent unreasonable risks, adopted regulations to update the scope of models covered by SB 1047, and established auditing standards.

SB 1047 would have required a comprehensive set of safety protocols prior to training a frontier model, including cybersecurity safeguards, the capability to execute a system-wide shutdown if the model proved dangerous, and reasonable measures to prevent critical harm. Before deployment, developers would have been required to assess whether their model could cause or materially enable critical harms, retain the results of such assessments, and make reasonable efforts to implement safeguards. The bill would have also prohibited the release of any model that posed an unreasonable risk or could enable critical harm.

Additionally, SB 1047 would have required developers to retain a third-party auditor to conduct independent assessments of their compliance with the bill. Records generated under SB 1047 would have been made available in redacted form to both the public and the AG, with the AG having the authority to request unredacted copies.

Beyond the Board and the bill's safety and transparency provisions, SB 1047 would have required computing clusters to implement procedures to evaluate whether customers intended to use their infrastructure to train a covered model. The bill also would have established, within GovOps, a consortium tasked with developing a framework for a public cloud computing cluster, CalCompute, to support the safe development and deployment of AI. SB 1047 also included whistleblower protections, allowing employees to report noncompliance to either the Labor Commissioner or the AG.

Lastly, SB 1047 would have imposed significant penalties on developers if their model caused death or bodily harm, damage to property, theft or misappropriation of property, or posed an imminent risk to public safety. For a first offense, developers could face penalties of up to 10% of the compute cost used to train the model, increasing to 30% for repeat offenses. Additionally, penalties for operators of computer clusters that violated the bill would start at $50,000 for a first offense and $100,000 for subsequent violations. The AG would also have been authorized to seek injunctive or declaratory relief, monetary or punitive damages, attorney's fees and costs, and any other form of relief deemed appropriate.

Ultimately, SB 1047 was vetoed by Governor Gavin Newsom. In his veto message, the Governor stated:

---

[34] Dan Hendryks, *Introduction to AI Safety, Ethics, and Society,* p. 12, https://drive.google.com/file/d/1uph559W-ASR4MEn6M_7Mb3lqQTapC_gZ/view?pli=1.

> By focusing only on the most expensive and large-scale models, SB 1047 establishes a regulatory framework that could give the public a false sense of security about controlling this fast-moving technology. Smaller, specialized models may emerge as equally or even more dangerous than the models targeted by SB 1047 – at the potential expense of curtailing the very innovation that fuels advancement in favor of the public good.
>
> Adaptability is critical as we race to regulate a technology still in its infancy. This will require a delicate balance. While well-intentioned, SB 1047 does not take into account whether an AI system is deployed in high-risk environments, involves critical decision-making or the use of sensitive data. Instead, the bill applies stringent standards to even the most basic functions – so long as a large system deploys it. I do not believe this is the best approach to protecting the public from real threats posed by the technology.
>
> Let me be clear – I agree with the author – we cannot afford to wait for a major catastrophe to occur before taking action to protect the public. California will not abandon its responsibility. Safety protocols must be adopted. Proactive guardrails should be implemented, and severe consequences for bad actors must be clear and enforceable. I do not agree, however, that to keep the public safe, we must settle for a solution that is not informed by an empirical trajectory analysis of AI systems and capabilities. Ultimately, any framework for effectively regulating AI needs to keep pace with the technology itself.
>
> To those who say there's no problem here to solve, or that California does not have a role in regulating potential national security implications of this technology, I disagree. A California-only approach may well be warranted – especially absent federal action by Congress – but it must be based on empirical evidence and science. The U.S. AI Safety Institute, under the National Institute of Science and Technology, is developing guidance on national security risks, informed by evidence-based approaches, to guard against demonstrable risks to public safety. Under an Executive Order I issued in September 2023, agencies within my Administration are performing risk analyses of the potential threats and vulnerabilities to California's critical infrastructure using AI. These are just a few examples of the many endeavors underway, led by experts, to inform policymakers on AI risk management practices that are rooted in science and fact. [. . .]

6) **Frontier Model Working Group and what this bill would do.** Following his veto of SB 1047, Governor Newsom commissioned the Joint California Policy Working Group on AI Frontier Models to prepare a report on the regulation of frontier models. The Working Group was led by Dr. Fei-Fei Li, Co-Director of the Stanford Institute for Human-Centered Artificial Intelligence; Dr. Mariano-Florentino Cuéllar, President of the Carnegie Endowment for International Peace; and Dr. Jennifer Tour Chayes, Dean of the UC Berkeley College of Computing, Data Science, and Society. In June 2025, the Working Group released their report, which highlighted the issues such as transparency, incident reporting, scoping, and independent evaluations. This bill incorporates some of the Working Group's recommendations to create a narrow framework to ensure transparency and promote safety among frontier model developers.

Scoping. A major question that must be addressed before implementing any transparency measure or incident reporting requirements is: What kinds of risks are especially concerning, and is there an evidentiary basis to believe that such harms could occur due to a large developer's frontier model? The Working Group recommends that:

> [P]olicymakers center their calculus around the marginal risk: Do foundation models present risks that go beyond previous levels of risks that society is accustomed to from prior technologies, such as risks from search engines?

To that end, this bill defines **"**catastrophic risk" as a foreseeable and material risk that a large developer's development, storage, use, or deployment of a foundation model will materially contribute to either:

- the death of, or serious injury to, more than 50 people; or
- more than one billion dollars ($1,000,000,000) in damage to rights in money or property.

Such harm must arise from a single incident in which a frontier model does any of following:

1. Provides expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon.
2. Engages in conduct with no meaningful human oversight, intervention, or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft, including theft by false pretense.
3. Evades the control of its large developer or user.

Each of these represents a capability that, prior to the advent of frontier models, would have required expert-level knowledge. For example, a search engine might direct someone to search for information about the most deadly pathogens or those most likely to cause a pandemic; however, a frontier model can synthesize that information and guide a user on how to manufacture a previously unknown pathogen with deadly capabilities. Similarly, while launching a large-scale cyberattack once required the acumen of a skilled computer scientist, a frontier model can not only write the underlying code for a virus or malware, but also autonomously identify backdoors and other exploitable vulnerabilities. Because of this ability to operate with minimal or no human prompting, frontier models have the potential to commit crimes, deceive users, or evade control in ways that previous technologies could not. A recent amendment clarifies that the loss of value of equity does not count as damage to or loss of property for the purposes of this bill.

Next, the question is: Who will be required to comply with this bill? Regarding scoping, the Working Group recommends:

> Since policy may have different regulatory intents and existing thresholds vary in their profiles of determination time, measurability, and external verifiability, we agree with Nelson et al. [90] that "a one-size-fits-all approach or a single threshold metric is inadequate for governance because different AI systems and their outputs present unique challenges and risks." To this end, we point to the European Union's AI Act, which designates models trained with 10^{25} FLOP as posing systemic risk as of March 2025 as the default criteria. However, the AI Act in Annex XIII affords the regulator flexibility to also consider alternative metrics, such as the number of parameters, size of the dataset, estimated cost or time of training, estimated energy consumption, benchmarks and evaluations of capabilities of the model, and whether the model has a high impact on the internal market due to its reach (either due to at least 10,000 registered business users or the number of registered end users). Further, to capture fast-moving scientific developments, the AI Act creates a scientific panel

> that is empowered to issue qualified alerts to identify models that may pose systemic risk even if they are not captured by predefined quantitative thresholds.
>
> Overall, we emphasize that irrespective of the combination of metrics deemed most appropriate in the present, policymakers should ensure that mechanisms exist not only to update specific quantitative values, given the rapid pace of technological and societal change in AI, but also to change the metrics altogether.[35]

This bill draws inspiration from SB 1047, EU AI Act, and the Working Group report. Recent amendments have also substantially updated the scoping of this bill. These amendments establish a two-tiered system of requirements under the bill. First, a "frontier model" is defined as a model trained using 10^{26} FLOP, a measure of computing power. Second, a "frontier developer" is defined as a person who has trained, or plans to train, a frontier model and who has access to the level of compute required to meet that threshold. The bill further defines a "large frontier developer" as a frontier developer whose gross revenue exceeded $500 million in the previous year. This tiered framework allows for the bill to place more stringent transparency obligations, outlined in the transparency section, on the better-resourced large frontier developers, while still ensuring that all frontier developers remain subject to baseline transparency and reporting requirements as well as whistleblower protections.

Recent amendments have also revised how this bill builds flexibility into its scope. Unlike SB 1047, which established the Board of Frontier Models, or earlier versions of this bill that granted the AG rulemaking authority to adjust the scope, these amendments instead vest advisory authority in the California Department of Technology (CDT). Beginning in 2027, CDT must provide recommendations to the Legislature on updating the definitions of "frontier model," "frontier developer," and "large frontier developer" to reflect technological advances. As model training becomes more efficient, the compute required to develop a model capable of catastrophic harm may decrease. In addition, as the Working Group report notes, compute alone may not remain the most appropriate proxy for catastrophic risk. Accordingly, this bill authorizes CDT to recommend updated definitions that may go beyond purely quantitative thresholds, such as raw compute or revenue, to incorporate criteria based on model capabilities or other relevant factors.

In making these recommendations, CDT must consider standards and guidance from other jurisdictions, including federal and international bodies, and engage in a stakeholder process that includes input from academics, industry representatives, the open-source community, and government entities. This process is intended to ensure that, if the Legislature updates the definitions, the resulting definitions provide clarity for developers regarding their obligations under the law.

Transparency. Having established who is subject to the bill, the legislation sets forth a transparency regime. These procedures are designed to provide insight into how large frontier developers manage, assess, and mitigate catastrophic risks. This approach aligns with the Working Group's recommendation to implement robust safety practices:

---

[35] Bommasani,and Singer et al.. "The California Report on Frontier AI Policy." The Joint California Policy Working Group on AI Frontier Models. June 17,2025. p 39.

> Transparency into the risks associated with foundation models, what mitigations are implemented to address risks, and how the two interrelate is the foundation for understanding how model developers manage risk. In turn, this information directly informs how other entities in the supply chain should modify or implement safety practices. In addition, transparency into the safety cases used to assess risk provides clarity into how developers justify decisions around model safety.[36]

This bill incorporates transparency requirements within a broader framework, termed the frontier AI framework (framework), which large frontier developers must draft, implement, and publish on their websites. The framework must include:

- *Defining and Assessing Thresholds*: An explanation of how the developer assesses catastrophic risks, including the capability thresholds the developer will use and whether those risks arise from misuse or model evasion.
- *Mitigation Strategies:* A disclosure of the measures used to mitigate catastrophic risks, how the developer evaluates their effectiveness, and whether third parties are involved in the assessment.
- *Cybersecurity Practices:* A summary of the cybersecurity safeguards in place to protect model weights from unauthorized access or modification.
- *Incident Response Plans*: A description of how the developer would respond to a critical incident involving their model, as well as how they manage risks arising from internal use of the model.

The framework serves as a core transparency mechanism, ensuring that large developers maintain a baseline standard of transparency for their processes. Recent amendments further require the large frontier developer to review and, if needed, update their framework at least once per year.

In tandem with the framework, the bill also requires both large frontier developers and frontier developers to submit transparency reports at the time of deploying a foundation model. The Working Group draws a parallel between these reports and the historical conduct of the tobacco industry, which concealed its knowledge that smoking causes lung cancer. In contrast, this bill seeks to prevent such obfuscation by mandating upfront disclosures about the potential risks and safety practices surrounding advanced AI models:

> The history of the tobacco industry reveals the importance of developing frameworks that promote transparency around companies' internal risk assessments and research findings. In the AI context, frontier AI labs possess the most holistic information about their models' capabilities and risks. Making this information accessible to policymakers and external experts can promote policy informed by a holistic understanding of the state-of-the-art of evidence produced by those closest to the technology, supporting informed oversight without stifling innovation.[37]

It is essential for decision-makers to understand the real, material harms that could arise from these models and to guide policy based on that knowledge. In the foundation model space, such

---

[36] *Id.* at p. 26.
[37] *Id.* at p. 19.

disclosures are typically provided at deployment in documents known as model cards. However, these model cards vary widely in detail and depth depending on the developer, which can create the false impression that some foundation models are inherently safer or better than others.

The recent amendments require large frontier developers to publish a transparency report before or at the time of deploying a foundation model. This report must include the results of any risk assessments, mitigation steps, and evaluations of their effectiveness as outlined in the developer's framework as well as the extent to which third parties were used in these evaluations. All frontier developers must publish a transparency report that is substantially narrower, requiring only the internet website of the frontier developer, a mechanism that enables a natural person to communicate with the frontier developer, the release date of the frontier model, the languages supported by the frontier model, the modalities of output supported by the frontier model, the intended uses of the frontier model, and any generally applicable restrictions or conditions on uses of the frontier model. The bill clarifies that the requirements of the transparency report may be met via the model system card.

Recent amendments also mandate that large frontier developers to transmit a summary report of catastrophic risk assessments of their frontier models to OES. This is particularly important because the most serious risks may emerge well before deployment. While transparency reports provide insight into risks associated with deployed models, they only report on models that have been released. As noted in the Working Group report:

> Sophisticated AI systems, when sufficiently capable, may develop deceptive behaviors to achieve their objectives, including circumventing oversight mechanisms designed to ensure their safety. Because these risks are unique to AI compared to other technologies, oversight is critical for external outputs as well as internal testing and safety controls. Policies that govern internal deployment are common for high-risk emerging technologies.[38]

Ultimately, this bill creates a transparency framework that will give some insight and scrutiny to the processes from initial training of a foundation model all the way to post deployment.

Adverse Event Reporting. A major component of understanding the impact of foundation models on society requires strong post deployment monitoring and accountability. The Working Group suggests:

> An adverse event reporting system that combines mandatory developer reporting with voluntary user reporting maximally grows the evidence base. A hybrid model of mandatory and voluntary reporting requirements in designing an adverse event reporting system can maximize the robust evidence base necessary for adverse event reporting systems to function properly. For example, a system could require mandatory reporting for AI model developers that operates in tandem with voluntary reporting for downstream users.[39]

The recent amendments incorporate this recommendation by tasking OES with creating a mechanism for critical incident reporting. The bill defines a **"**critical incident" as any of the following:

---

[38] *Id.* at p. 21.
[39] *Id.*at p. 35.

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

Under this mechanism, a frontier developer or a member of the public may report a critical safety incident to OES. Reports must include the date of the event, an explanation of how it qualifies as a critical incident, a detailed description of the event, and whether the incident was associated with internal use of a frontier model. Frontier developers are required to report any critical safety incident within 15 days of discovering it. OES must review all reports submitted by frontier developers but may choose whether or not to review reports made by the public. This reporting mechanism aims to establish a system in which potential harms are identified and mitigated before escalating into catastrophes, while also fostering greater cooperation between government and the private sector to address such risks.

The bill further mandates that frontier developers immediately notify the appropriate law enforcement authority in the event of a critical incident, such as the detection that their model helped develop a bioweapon. These authorities are better equipped to respond swiftly and effectively in ways OES may not be. After alerting law enforcement, large developers would then still have 15 days to report the critical incident to OES. Furthermore, the bill enables developers to revise their incident report at a future date in the event more information is learned about the incident.

The bill further requires OES to publish an anonymized and aggregated summary of all critical incident and whistleblower reports. These public summaries **will not reveal trade secrets, the identities of reporters, or which frontier developer the report concerns.** Additionally, the amendments grant OES discretion to share reports with the Governor, relevant state departments, or the Legislature when warranted. This approach will help bridge the knowledge gap between regulators and industry, foster greater cooperation, and ensure that decisionmakers are informed about the current state of advanced technologies and the risks they may pose. Reports of catastrophic risk assessments from internal use are still shielded from public disclosure.

Lastly, this bill gives OES the ability to adopt regulations that designate one or more federal laws, regulations, or guidance documents as being in compliant with this bill. These regulations must ensure that any other law deemed to meet the standards of this bill must be equivalent or stricter than this bill, and intended to assess and mitigate catastrophic risk. This may lay the framework for a national standard for adverse event reporting.

CalCompute. This bill, like SB 1047, also establishes a consortium within the GovOps to develop a framework for creating a public cloud computing cluster known as **"CalCompute."** This initiative responds to the fact that academic institutions currently lack sufficient computing power to conduct research at the scale of large developers. This creates a resource and research gap, where the academic institutions, typically responsible for studying the safe and effective use of new technologies, are unable to keep pace with advancements at the AI frontier.

AI has the potential to transform our economy and power new industries; however, this transformation can only be fully and equitably realized with public support. The establishment of CalCompute aims to advance that goal by ensuring academic institutions have the necessary resources to conduct essential research on foundation models that will inform and protect the public.

Specifically, the consortium will develop a framework for CalCompute that promotes the safe, ethical, equitable, and sustainable use of AI. The framework development will include a report analyzing the state's current cloud computing infrastructure, the costs of building and maintaining CalCompute, and the state's technology workforce. The report will also offer recommendations for equitable pathways to strengthen the workforce and outline CalCompute's role in supporting these efforts.

Furthermore, the report must include recommendations for CalCompute's governance and operation, usage parameters, and how its creation and ongoing management can prioritize the employment of the current public sector workforce. The bill requires CalCompute to feature a fully owned and hosted cloud platform, staffed with the necessary human expertise to operate, maintain, support, train, and facilitate its use.

The consortium must prioritize locating CalCompute within the University of California system. If established there, CalCompute may also accept private donations. The consortium will consist of four representatives from the University of California and other public and private academic research institutions and national laboratories, along with four technology and AI experts appointed by the Secretary of Government Operations to provide technical assistance. The Speaker of the Assembly will appoint three representatives from impacted workforce labor organizations, and the Senate Rules Committee will appoint three representatives from stakeholder groups with relevant expertise and experience.

Whistleblower Protections. Another aspect of transparency addressed by the Working Group Report is whistleblower protections. The report states:

> Different existing whistleblower protections tend to apply when two conditions are satisfied: (i) The whistleblower is blowing the whistle on appropriate topics; and (ii) the whistleblower follows established reporting protocols. In terms of the topics that qualify for protection, prior work, based on a survey of existing whistleblower protections across multiple jurisdictions (e.g., the United States at the federal level, the European Union), finds that many existing protections across different sectors share a focus on violations of the law. However, actions that may clearly pose a risk and violate company policies (e.g., releasing a model without following the protocol laid out in a company's safety policy) may not violate any existing laws. Therefore, policymakers may consider protections that cover a broader range of, activities, which may draw upon notions of "good faith" reporting on risks found in other domains such as cybersecurity.[40]

The bill takes these recommendations into account. The provisions and merits of the whistleblower section fall within the jurisdiction of the Assembly Judiciary Committee, which has thoroughly analyzed this part of the bill. It should be noted that the bill expands the

---

[40] *Id.*at p. 29.

whistleblower protections to include disclosures concerning falsehoods or misrepresentations in the large developers' framework or transparency reports. Recent amendments have removed contractors from the definition of "covered employees"; as a result, third parties, such as red-teamers or auditors, are not protected by the whistleblower protections. Lastly, similarly to the OES reports on critical safety incidences, this requires the AG to publish an anonymized and aggregated summary of all whistleblower reports.

Enforcement. The recent amendments have revised how this bill is enforced. Previously, the bill, like SB 1047, provided for independent audits for compliance with the bill, starting in 2030. This provision, which aligned with the Working Group's emphasis on independent evaluation frontier model safety protocols, was removed in recent amendments. Additionally, the bill had a multitiered approach in which the amount of the violation scaled with the violation. Now, the bill states that the a large frontier developer who violates this bill by failing to publish or transmit a compliant document that is required, fails to report an incident, or fails to comply with its own framework shall be subject to a civil penalty subject in an amount dependent upon the severity of the violation that does not exceed one million dollars per violation. Notably this is a much more lenient enforcement mechanism and penalty than those instituted in SB 1047.

7) **Comparison to the RAISE Act.** This bill draws comparisons to the Responsible AI Safety and Education Act (RAISE Act), which recently passed both houses of the New York Legislature and now awaits a decision from Governor Kathy Hochul.[41] Like this bill, the RAISE Act requires a safety framework detailing the developer's risk mitigation practices, enshrines whistleblower protections for employees and contractors, and mandates critical incident reporting. Both bills also similarly define the types of risks and critical incidents that must be addressed. However, the two differ significantly in the timeline for reporting such incidents: the RAISE Act requires reporting within 72 hours of becoming aware of a critical incident, while SB 53 allows for 15 days, with an exception for imminent dangers which must be reported to law enforcement with 24 hours.

Other key differences include SB 53's requirement for a transparency report for deployed models and the establishment of internal incident reporting mechanisms, both of which were recommended in the Working Group Report. Additionally, SB 53 grants the CDT the authority to issue recommendations to the Legislature on the definition of a "large frontier developer", "frontier developer" and "frontier model", a built-in opportunity for flexibility the RAISE Act does not include. While liability under SB 53 is capped to a maximum of one million dollars, the RAISE Act imposes civil penalties of up to $10 million for first-time violations of its transparency requirements and up to $30 million for repeat offenses, as well as $10,000 per violation of its whistleblower provisions.

ARGUMENTS IN SUPPORT: Anthropic, writes in support:

> As you know, SB 53 would, for the first time, govern powerful AI systems built by frontier AI developers like Anthropic. We've long advocated for thoughtful AI regulation and our support for this bill comes after careful consideration of the lessons learned from California's previous attempt at AI regulation (SB 1047). While we believe that frontier AI safety is

[41] RAISE Act can be found at https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/A.

ideally addressed at the federal level instead of a patchwork of state regulations, powerful AI advancements won't wait for consensus in Washington.

The measure is also in keeping with direction from Governor Newsom and his Joint California Policy Working Group. The working group endorsed an approach of 'trust but verify', and SB 53 implements this principle through disclosure requirements rather than the prescriptive technical mandates that plagued last year's efforts.

SB 53 would require large companies developing the most powerful AI systems to:

- Develop and publish safety frameworks, which describe how they manage, assess, and mitigate catastrophic risks—risks that could foreseeably and materially contribute to a mass casualty incident or substantial monetary damages.

- Release public transparency reports summarizing their catastrophic risk assessments and the steps taken to fulfill their respective frameworks before deploying powerful new models.

- Report critical safety incidents to the state within 15 days, and even confidentially disclose summaries of any assessments of the potential for catastrophic risk from the use of internally-deployed models.

- Provide clear whistleblower protections that cover violations of these requirements as well as substantial dangers to public health/safety from catastrophic risk.

- Be publicly accountable for the commitments made in their frameworks or face monetary penalties.

These requirements would formalize practices that Anthropic and many other frontier AI companies already follow. At Anthropic, we publish our Responsible Scaling Policy, detailing how we evaluate and mitigate risks as our models become more capable. We release comprehensive system cards that document model capabilities and limitations. Other frontier labs (Google DeepMind, OpenAI, Microsoft) have adopted similar approaches while vigorously competing at the frontier. Now all covered models will be legally held to this standard. The bill also appropriately focuses on large companies developing the most powerful AI systems, while providing exemptions for smaller companies that are less likely to develop powerful models and should not bear unnecessary regulatory burdens. Of course, no major piece of legislation like SB 53 is perfect, nor do we expect it to be. But what is clear is that SB 53's transparency requirements will have an important impact on frontier AI safety. Without it, labs with increasingly powerful models could face growing incentives to dial back their own safety and disclosure programs in order to compete. But with SB 53, developers can compete while ensuring they remain transparent about AI capabilities that pose risks to public safety, creating a level playing field.

The question before us all isn't whether we need AI governance—it's whether we'll develop it thoughtfully today or reactively tomorrow. SB 53 offers a solid path toward the former. We commend Senator Wiener and Governor Newsom for their leadership on responsible frontier AI governance, and we encourage the California Legislature to pass SB 53.

ARGUMENTS IN OPPOSITION: In an oppose-unless-amended position, CalChamber, Computer & Communications Industry Association, and TechNet jointly write:

> [. . .]
>
> We share your goal of ensuring the safe and responsible development of AI and appreciate efforts made in recent amendments to find common ground on how California should approach artificial intelligence models and we appreciate improvements made to the bill over the last several weeks. That being said, there are some issues of concern that remain and wish to flag certain other areas where the bill could be better aligned with the final findings of Governor Newsom's Joint California Policy Working Group on AI Frontier Models, which arose out of his veto of SB 1047 (2024).
>
> **SB 53 should focus on model risk, not developer size—to fully address concerns about powerful models capable of catastrophic risk**
>
> We are concerned about the bill's focus on "large developers" to the exclusion of other developers of models with advanced capabilities that pose risks of catastrophic harm. As amended September 5th, **SB 53** now focuses on models that have a computational threshold of 10^26 floating point operations (or "FLOPs") but only if those models are developed by entities with at least $500m in annual revenues.
>
> Consistent with our position in SB 1047, we maintain that small entities can develop hugely influential and potentially risky models with similar capabilities to the models developed by "large developers", as demonstrated by the Chinese company DeepSeek. As noted above, upon vetoing SB 1047, the Governor commissioned experts in the field to form the Joint California Working Group on AI Frontier Models, which has validated such concerns in their Final Reports, finding that small companies may create powerful models that pose safety risks. By excluding such models here, the bill fails to adequately address the very real risks posed by small but malicious models and imposes significant costs on innovating performant but responsible ones. The Governor's Joint California Policy Working Group on AI Frontier Models cautions against developer-level thresholds stating:
>
> > Generic developer-level thresholds seem to be generally undesirable given the current AI landscape. Since many small entities can develop hugely influential and potentially risky foundation models, as demonstrated by the Chinese company DeepSeek, the use of thresholds based on developer-level properties may inadvertently ignore key players. […] At the same time, these approaches may bring into scope massive, established companies in other industries that are simply exploring the use of AI since thresholds based on properties of companies may not distinguish between the entire business and the AI-specific subset. Therefore, we caution against the use of customary developer-level metrics that do not consider the specifics of the AI industry and its associated technology.[42]

[42] Final Report at p.

SB 53 should make clear that the AI ecosystem includes multiple actors including downstream developers

SB 53 does not account for the complexity of the AI value chain. Models are routinely adapted and fine-tuned by downstream developers in ways that could potentially increase risk. The bill should make clear that a frontier developer's obligations do not extend to models that have been substantially modified by unaffiliated parties, otherwise accountability will be muddled and innovation chilled. We note that whereas the Governor's Work Group report recognized the full AI ecosystem value chain, **SB 53** still needs to fully recognize the roles of not just the original developer of a foundational model but also of those unaffiliated third parties who may modify and/or build on top of a foundation model. The bill should clarify these provisions to reflect the realities of the ecosystem, including downstream developers and open-source models.

SB 53 still raises concerns about protecting trade secrets and sensitive information, including matters of cybersecurity and national security.

We appreciate that amendments were made to change the level of detail required of the AI Safety Framework and changing summaries for transparency reports. However, **SB 53** now requires a large developer only to transmit to the California Office of Emergency Services (CalOES) a summary of any assessment of catastrophic risk resulting from internal use of its frontier models every three months. Not only is this cadence of reporting unnecessary, CalOES will need to take serious steps to protect this information from being accessed by cybercriminals, foreign adversaries, or bad actors. Without ironclad safeguards, these transparency requirements could unintentionally make us less safe. The Joint California Policy Working Group on AI Frontier Models warns against this level of disclosure.

> General details about risks of foundation models can be made public without undermining security, especially if these risks have been demonstrated in other foundation models or AI technologies. Specific details about concrete vulnerabilities should be disclosed carefully, with advanced notice to actors in the supply chain who are able to remediate them prior to broader disclosure.[43]

Requiring developers to justify redactions is less effective than not requiring developers to disclose any information that would include trade secrets, cybersecurity information, or other confidential or proprietary information.

SB 53 unnecessarily re-writes California Whistleblower law for just one industry

As amended, SB 53 rewrites California's already robust whistleblower protections for just one industry. Creating a special, one-off standard for a single sector not only sets a poor precedent but also risks confusion and inconsistency across industries. Current law covers whistleblowing activities associated with AI safety because there is a robust body of existing law that governs whistleblower protection covering employees who report violations of state/federal laws, rules, or regulations. These laws are intentionally tied to actions that are illegal so there are clear lines of what is considered applicable and understood who gets

[43] *Id.* at 30.

protection when reporting. These protections cover activities associated with AI without creating unnecessary and confusing new processes in state law.

For example, Labor Code Section 6310 already protects whistleblowers who report unsafe working conditions or work practices. Similarly, federal laws such as the Sarbanes-Oxley Act protect employees who report safety violations or substantial and specific dangers to public health or safety. A brightline threshold is needed for what activity is covered so it is clear when a developer's activities should be reported. For instance, in the field of research and development, innovations are being experimented with in novel contexts where there may be significant disagreement on what actions constitute risk. Thus, the bill mandates that there be an allegation of "*specific and substantial danger to public health or safety resulting from a catastrophic risk*," the inherently subjective nature of these terms leaves room for differing interpretations as to what does or does not meet the threshold.

SB 53 requires steep penalties that are disproportionate for technical errors, inflexible incident reporting requirements, and no right to cure

As amended, **SB 53** imposes a $1 million fine for a possible paperwork error which is excessive and risks punishing good-faith developers for technical mistakes rather than deterring real harm. Penalties should be fair, targeted, and proportionate. As we pointed out in our July 12th letter, **SB 53** requires incident reporting within 15 days but does not provide flexibility for an investigation timeline. Even if 15 days is a reasonable reporting period, requirements should be flexible because all facts may not be known within 15 days of discovery. With respect to enforcement, we again state our view that the bill should grant businesses at least a 60 day right to cure, to ensure that law focuses on compliance and not punishment. In addition, given the highly detailed requirements of the bill as drafted, we think enforcement efforts should be focused on material failures to comply rather than also covering technical paperwork errors.

While we understand your focus on this issue and appreciate the recent amendments have made meaningful improvements to the prior version of the bill, given the immense promise of this technology, we believe that the bill would benefit from a focus on a risk-based framework for all frontier models, additional clarity in responsibilities among actors in the AI value chain, additional safeguards for trade secrets and security, and reasonable timelines, penalties, and enforcement provisions. [. . .]

REGISTERED SUPPORT / OPPOSITION:

Support

Ai for Animals
Ai Futures Project
Ai Lab Watch
Ai Policy Tracker
All Girls Allowed
Anthropic Pbc
Apart Research
Association for Long Term Existence and Resilience (ALTER)

Berkeley Existential Risk Initiative (BERI)
California Federation of Labor Unions, Afl-cio
Center for Ai and Digital Policy
Center for Ai Policy
Center for Digital Democracy
Center for Human-compatible Ai
Center for Youth and Ai
Children's Advocacy Institute, University of San Diego School of Law
Common Sense Media
Depict.ai
Design It for US
District Council of Iron Workers of the State of California and Vicinity
Earningsstream LLC
Economic Security California Action
Elicit
Encode
Encode Ai Corporation
Encode Justice
Eon Systems
Existential Risk Observatory
Frontlines Foundation
Future of Life Institute
Indivisible California Statestrong
InnovateEDU
Momentum
Mothers Against Media Addiction
Nonlinear
Noso November
Oakland Privacy
Parents Television and Media Council
Parents Together Action
Public Interest Privacy Center
Redwood Research
Rights4girls
Scorecard
Secure Ai Future
Secure Ai Project
SEIU Califonia
Tech Oversight California
Techequity Action
The Brandes Lab At Nyu
The Midas Project
The Signals Network
Transparency Coalition.ai
Trevi Digital Assets Fund
University of California
Young People's Alliance
Youth Power Project

Opposition

Business Software Association
Chamber of Progress
Consumer Technology Association
Los Angeles County Business Federation (BIZFED) (UNREG)
Silicon Valley Leadership Group

Oppose Unless Amended

California Chamber of Commerce
Computer and Communications Industry Association
Insights Association
Technet

Analysis Prepared by:John Bennett / P. & C.P. / (916) 319-2200