Date of Hearing: July 16, 2025
Fiscal: Yes

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
SB 53 (Wiener) – As Amended July 8, 2025

SENATE VOTE: 37-0

PROPOSED AMENDMENTS

SUBJECT: CalCompute: foundation models: whistleblowers

SYNOPSIS

"Frontier" or "foundation" models are the largest and most powerful AI systems in development. These models are highly generalizable and have the potential to drive breakthroughs in science and medicine, streamline complex processes, and strengthen the economy. However, they also pose serious catastrophic risks due to their immense capabilities. A frontier model could help cure disease or design the next pandemic. It could automate bureaucratic functions or become autonomous and disrupt critical infrastructure.

Critics of AI safety argue that the evidence base is still too limited to justify regulation, and that prematurely imposing safeguards could stifle innovation. Responding to this dilemma, the International AI Safety Report, led by one of the "godfathers of AI," Yoshua Bengio, has warned:

> *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.*

In the 2024 legislative session, SB 1047 (Wiener) sought to address concerns surrounding frontier models by establishing a regulatory framework intended to prevent catastrophic harms. The bill would have required frontier model developers to create and implement comprehensive safety and security protocols before initiating training, to implement shutdown capabilities, and to perform risk assessments on models and implement reasonable safeguards, subject to third-party auditing, before releasing the models. The bill also would have prohibited releasing or using models that pose an unreasonable risk of catastrophic harms. Finally, SB 1047 would have created a new state agency – the Board of Frontier Models – to oversee the development of these models.

In vetoing the bill, Governor Gavin Newsom acknowledged the risks but emphasized the importance of addressing the evidence dilemma:

> *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.

Following the veto, the Governor convened the Joint California Policy Working Group on AI Frontier Models, which used empirical research, historical case studies, and modeling 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 recommendations of the Working Group Report. Narrower than its predecessor, SB 53 takes a light-touch approach that focuses on transparency as the means of ensuring safety and accountability for the largest developers of the most powerful models. Large developers who harness an extraordinarily high amount of compute power must create, implement, and publish both a safety and security protocol and a transparency report for each released model. The bill does not prescribe any particular standards for these plans: it simply requires developers to explain whether and how they assess, mitigate, and manage catastrophic risks – those that would result in more than 100 deaths or $1 billion in damage.

The bill also establishes a critical incident reporting mechanism, administered by the Attorney General (AG), to ensure that severe or high-risk events are tracked and addressed in a timely manner. Incidents must be reported within 15 days. The AG is further authorized to update the definition of a "large developer" through rulemaking to ensure that the bill remains responsive to technological advancements. Additionally, the bill establishes a consortium within the Government Operations Agency to create a public computing cluster, known as CalCompute, to support AI research and safety testing. The bill also provides whistleblower protections for employees and contractors of large developers who report risks or noncompliance. The AG is authorized to enforce the bill by seeking an unspecified civil penalty.

The bill is sponsored by Encode Justice, Secure AI Project, and Economic Security California Action and has received support from a broad coalition of technology advocacy groups, including TechEquity Action, Secure AI Future, and Common Sense Media. It is opposed by the Silicon Valley Leadership Group and the Chamber of Progress. The California Chamber of Commerce alongside a coalition of technology trade organization have taken an opposed unless amended position. The bill previously passed the Assembly Judiciary Committee by a 12-0 vote.

Committee amendments outlined in Comment #8 strengthen the bill's transparency and accountability framework by incorporating additional recommendations from the Working Group Report. First, the Committee amendments require developers to include capability thresholds in both their safety and security protocols and transparency reports. Developers must disclose when their models exceed these thresholds and must document any mitigation measures taken in response. Second, to avoid over-burdening smaller startups, the amendments add to the definition of "large developer" an annual revenue threshold of $100 million, in addition to the current compute threshold. Third, the amendments require developers to disclose <u>whether</u> they retain the ability to shut down a model under their control in the event of a critical incident. This is a disclosure requirement only; unlike SB 1047, this bill does not mandate that a developer maintain the technical ability to shut down a model. Fourth, if a critical incident presents an imminent threat, the amendments require the developer to report the incident within 24 hours to the appropriate law enforcement authority. Fifth, the amendments direct the AG to publish an annual, anonymized, and aggregated summary of all critical incident and whistleblower reports. Sixth, the amendments implement the Working Group's recommendations to use third-party

assessments by requiring developers to undergo annual third-party audits to assess compliance with their own safety protocols. To give time for the nascent industry of AI auditors to grow to meet this demand, this requirement does not kick in until 2030. Finally, the Committee amendments flesh out the enforcement provisions.

THIS BILL:

1) Defines the following 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 large developer's development, storage, use, or deployment of a foundation model will materially contribute to the death of, or serious injury to, more than 100 people or more than one billion dollars ($1,000,000,000) in damage to rights in money or property arising from a single incident, scheme, or course of conduct involving a dangerous capability.

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

        i. Unauthorized access to, modification of, or exfiltration of, the model weights of a foundation model.

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

        iii. Loss of control of a foundation model causing death, bodily injury, or damage to rights in money or property.

        iv. A foundation model that uses deceptive techniques against the large developer to subvert the controls or monitoring of its large developer outside of the context of an evaluation designed to elicit this behavior.

    d. "Dangerous capability" to mean the capacity of a foundation model to do any of the following:

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

        ii. Conduct or assist in a cyberattack.

        iii. Engage in conduct, with limited human intervention, that would, if committed by a human, constitute the crime of extortion, theft, including theft by deception, or a serious or violent felony.

        iv. Evade the control of its large developer or user.

    e. "Deploy" to mean to make a foundation model available to a third party for use, modification, copying, or combination with other software. "Deploy" does not

include making a foundation model available to a third party for the primary purpose of developing or evaluating the foundation model.

f. "Large developer" to mean either of the following:

i. Before January 1, 2027, "large developer" to mean a person who has trained, or initiated the training of, at least one foundation model using a quantity of computing power greater than 10^26 integer or floating-point operations.

ii. Beginning on and after January 1, 2027, "large developer" has the meaning defined by a regulation adopted by the Attorney General. Clarifies that if the Attorney General does not adopt a regulation described in 7b by January 1, 2027, the definition in 7a will be operative until the regulation is adopted.

g. "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.

h. "Safety and security protocol" to mean the documented technical and organizational protocols to manage, assess, and mitigate catastrophic risks.

2) Requires a large developer to write, implement, and clearly and conspicuously publish on its internet website a safety and security protocol that describes in specific detail all of the following:

a. How, if at all, the large developer excludes certain foundation models from being covered by its safety and security protocol because those foundation models do not pose material catastrophic risks.

b. The testing procedures that the large developer uses to assess catastrophic risks from its foundation models, including risk resulting from malfunctions, misuse, and foundation models evading the control of the large developer or user.

c. The mitigations that a large developer takes to reduce a catastrophic risk and how the large developer assesses the effectiveness of those mitigations.

d. The degree to which the large developer's assessments of catastrophic risk and the effectiveness of catastrophic risk mitigations are reproducible by external entities.

e. The extent to which, and how, a large developer will use third parties to assess catastrophic risks and the effectiveness of mitigations of catastrophic risk.

f. The large developer's cybersecurity practices and how the large developer secures unreleased model weights from unauthorized modification or transfer by internal or external parties.

g. To the extent that the foundation model is controlled by the large developer, the procedures the large developer will use to monitor critical safety incidents and the steps that a large developer would take to respond to a critical safety incident, including, but not limited to, who the large developer will notify and the timeline on which the large developer would take these steps.

h. The testing procedures that the large developer will use to assess and manage a catastrophic risk resulting from the internal use of its foundation models, including risks resulting from a foundation model circumventing oversight mechanisms, and the schedule, specified in days, by which the large developer will report these assessments pursuant to 5.

i. How the developer determines when its foundation models are substantially modified enough to conduct additional assessments and publish a transparency report pursuant to 4.

3) Requires that if a large developer makes a material modification to its safety and security protocol, the large developer must clearly and conspicuously publish the modified protocol and a justification for that modification within 30 days.

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

a. The results of any risk assessment, the steps taken to address any identified risks, and the results of any risk mitigation assessment conducted by the large developer pursuant to its safety and security protocol.

b. The results of any risk assessment, the steps taken to address any identified risks, and the results of any risk mitigation assessment conducted pursuant to a large developer's safety and security protocol that is conducted by a third party that contracts with a large developer.

c. The time and extent of predeployment access provided to any third party described in subparagraph 4b, whether or not the third party was independent, and the nature of any constraints the large developer placed on the assessment or on the third party's ability to disclose information about its assessment to the public or to government officials.

d. If the deployment would pose a catastrophic risk, the reasoning behind the large developer's decision to deploy the foundation model, the process by which the large developer arrived at that decision, and any limitations in the assessments that the large developer used to make that decision.

e. Clarifies that a large developer may reuse an answer previously provided if the rationale in question has not materially changed for the new deployment.

5) Requires a large developer to clearly and conspicuously publish on its internet website any assessment of catastrophic risk resulting from internal use of its foundation models pursuant to the schedule the developer specifies in its safety and security protocol.

6) Prohibits a large developer from making a materially false or misleading statement about catastrophic risk from its foundation models or its management of catastrophic risk.

7) Permits a large developer that publishes required documents to make redactions to those documents that are necessary to protect the large developer's trade secrets, the large developer's cybersecurity, public safety, or the national security of the United States or to comply with any federal or state law.

8) Requires that if a large developer redacts information in a safety and security protocol, transparency report, or internal use reports, the large 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 retain the unredacted information for five years.

9) Requires the Attorney General to establish a mechanism to be used by a large 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.

10) Requires a large developer to report any critical safety incident pertaining to one or more of its foundation models to the Attorney General within 15 days of discovering the critical safety incident.

11) Requires the Attorney General to review critical safety incident reports submitted by large developers and permits the Attorney General to review reports submitted by members of the public.

12) States that a violation of this bill is subject to an unspecified civil penalty.

13) Requires that the civil penalty be assessed in a civil action brought only by the Attorney General.

14) Permits the Attorney General to adopt regulations to update the definition of a "large developer" for the purposes of this bill to ensure that it accurately reflects technological developments, scientific literature, and widely accepted national and international standards and applies to well-resourced large developers at the frontier of artificial intelligence development on or before January 1, 2027, and annually thereafter,.

15) Requires the Attorney General, in developing regulations pursuant to this section, to take into account all of the following:

    a. Similar thresholds used in international standards or federal law, guidance, or regulations for the management of catastrophic risk.

b. Input from stakeholders, including academics, industry, the open-source community, and governmental entities.

c. The extent to which a person will be able to determine, before beginning to train or deploy a foundation model, whether that person will be subject to the regulations as a large developer with an aim toward allowing earlier determinations if possible.

d. The complexity of determining whether a person is covered, with an aim toward allowing simpler determinations if possible.

e. The external verifiability of determining whether a person is covered, with an aim toward definitions that are verifiable by parties other than the large developer.

16) Requires that if the Attorney General determines that less well-resourced developers, or developers significantly behind the frontier of artificial intelligence, may create substantial catastrophic risk, the Attorney General must promptly submit a report to the Legislature with a proposal for managing this source of catastrophic risk but prohibits the Attorney General from including those developers within the definition of "large developer" without authorization in subsequently enacted legislation.

17) Establishes within the Government Operations Agency a consortium to develop a framework for the creation of a public cloud computing cluster to be known as "CalCompute."

18) Requires the consortium to develop a framework for the creation of 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.

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

20) 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.

21) Requires the consortium to operate in accordance with all relevant labor and workforce laws and standards.

22) Requires Government Operations Agency to submit a report from the consortium to the Legislature with the framework developed by this bill for the creation and operation of CalCompute on or before January 1, 2027.

23) Requires the report required by 22 to include all of the following elements:

a. A landscape analysis of California's current public, private, and nonprofit cloud computing platform infrastructure.

b. An analysis of the cost to the state to build and maintain CalCompute and recommendations for potential funding sources.

c. Recommendations for the governance structure and ongoing operation of CalCompute.

d. Recommendations for the parameters for use of CalCompute, including, but not limited to, a process for determining which users and projects will be supported by CalCompute.

e. An analysis of the state's technology workforce and recommendations for equitable pathways to strengthen the workforce, including the role of CalCompute.

f. A detailed description of any proposed partnerships, contracts, or licensing agreements with nongovernmental entities, including, but not limited to, technology-based companies, that demonstrates compliance with the requirements of 19 and 20.

g. Recommendations regarding how the creation and ongoing management of CalCompute can prioritize the use of the current public sector workforce.

24) 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.

25) States that the members of the consortium serve without compensation, but will be reimbursed for all necessary expenses actually incurred in the performance of their duties.

26) Requires the consortium to be dissolved upon submission of the report to the Legislature.

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

28) Requires that CalCompute become operative only upon an appropriation.

29) Prohibits a large developer from making, adopting, enforcing, or entering into a rule, regulation, policy, or contract that prevents an employee from disclosing, or retaliates against

an employee for disclosing, information to the Attorney General, a federal authority, a person with authority over the employee, or another employee who has authority to investigate, discover, or correct the reported issue, if the employee has reasonable cause to believe that the information discloses either of the following:

a. The large developer's activities pose a catastrophic risk.

b. The large developer has violated the transparency elements of this bill.

30) Prohibits a large developer from entering into a contract that prevents an employee from making a disclosure protected under Labor Code section 1102.5.

31) Prohibits a large developer from making, adopting, enforcing, or entering into a rule, regulation, policy, or contract that would prevent an organization or entity that provides goods or services to the large developer related to the assessment, management, or addressing of catastrophic risk, or an employee of that organization or entity, from disclosing information to the Attorney General, a federal authority, or the developer if the organization, entity, or individual has reasonable cause to believe that the information discloses either of the following:

a. The large developer's activities pose a catastrophic risk.

b. The large developer has violated the transparency elements of this bill.

32) Permits an employee to use the hotline to make reports.

33) Requires a large developer to provide a clear notice to all employees of their rights and responsibilities under this bill, including by doing either of the following:

a. At all times posting and displaying within any workplace maintained by the large developer a notice to all employees of their rights under the bill, ensuring that any new employee receives equivalent notice, and ensuring that any employee who works remotely periodically receives an equivalent notice.

b. At least once each year, providing written notice to each employee of the employee's rights under the bill and ensuring that the notice is received and acknowledged by all of those employees.

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

35) Requires the disclosures and responses of the process required by this bill to be shared with officers and directors of the large developer at least once each quarter.

36) Clarifies that if an employee has alleged wrongdoing by an officer or director of the large developer in a disclosure or response, the disclosures described in 36 are prohibited with respect to that officer or director.

37) Establishes that the court is authorized to award reasonable attorney's fees to a plaintiff who brings a successful action for a violation of this section.

38) Requires the large developer to have the burden of proof to demonstrate by clear and convincing evidence that the alleged action would have occurred for legitimate, independent reasons even if the employee had not engaged in activities protected by this bill in a civil action brought pursuant to this section, once it has been demonstrated by a preponderance of the evidence that an activity proscribed by this section was a contributing factor in the alleged prohibited action against the employee.

39) Permits an employee to petition the superior court in any county wherein the violation in question is alleged to have occurred, or wherein the person resides or transacts business, for appropriate temporary or preliminary injunctive relief, in a civil action or administrative proceeding.

40) Requires the petitioner to cause notice thereof to be served upon the person, and thereupon the court to have jurisdiction to grant temporary injunctive relief as the court deems just and proper upon the filing of the petition for injunctive relief,.

41) Requires the court to consider the chilling effect on other employees asserting their rights under this bill in determining whether temporary injunctive relief is just and proper in addition to any harm resulting directly from a violation of this section,.

42) Requires that appropriate injunctive relief be issued on a showing that reasonable cause exists to believe a violation has occurred.

43) Requires an order authorizing temporary injunctive relief to remain in effect until an administrative or judicial determination or citation has been issued, or until the completion of a review, whichever is longer, or at a certain time set by the court. Permits, thereafter, that a preliminary or permanent injunction to be issued if it is shown to be just and proper. Prohibits any temporary injunctive relief from prohibiting a large developer from disciplining or terminating an employee for conduct that is unrelated to the claim of the retaliation.

44) Prohibits injunctive relief granted pursuant to this bill from not being stayed pending appeal.

EXISTING LAW:

1) Establishes the Government Operations Agency. (Gov. Code § 12800.)

2) Establishes the Department of Technology within the Government Operations Agency. (Gov. Code § 12803.2.)

3) Charges the Department of Technology 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 Section 1102.5.)

5) Requires the office of the Attorney General 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 Attorney General 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 Attorney General or appropriate government agency shall 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 Section 1102.7.)

COMMENTS:

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

In 2024, as part of his veto of Senate Bill 1047 (Wiener), Governor Newsom's Joint California Working Group on AI Frontier Models was established – a group of top experts tasked with charting a course forward on AI policy for the developers of the most advanced AI systems. Their final report, released in June 2025, emphasized the growing evidence for risk of severe harm, such as "AI-enabled hacking or biological attacks, and loss of control" and argued "California has a unique opportunity to continue supporting developments in frontier AI while addressing substantial risks that could have far-reaching consequences for the state and beyond."

Drawing recommendations from Governor Newsom's working group report, Senate Bill 53 requires covered developers to write, implement, and publish their safety and security protocol in redacted form to protect intellectual property. It would also require covered developers to report certain, carefully defined critical safety incidents to the Attorney General and would allow members of the public to report incidents.

SB 53 only applies to AI companies that have trained a model with 10^{26} floating point operations (FLOPs), a measure of computational power. These companies are spending hundreds of millions of dollars to train the most advanced AI models. As recommended by the Report, SB 53 also authorizes the Attorney General to adjust the scoping of the bill in the future to keep up with technological developments, but only focuses on well-resourced AI companies at the frontier of AI development.

Senate Bill 53 strengthens whistleblower protections for employees of frontier artificial intelligence laboratory companies whose activities pose a catastrophic risk. SB 53 also establishes a consortium to help create CalCompute: a public AI research cluster that will provide startups and researchers with access to the resources needed to develop large-scale AI systems.

In doing this, and building on Governor Newson's working group report, SB 53 allows California to continue to maintain its leadership in the AI development ecosystem 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

---

[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/.

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] 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.

---

[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.

- 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, 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

---

[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.

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, 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 certain 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]

---

[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.
[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.

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 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 blackmailed 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,

---

[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./.
[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.

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 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} FLOP and costing over $100 million. This board, known as the Board of Frontier Models, would have been housed within the Government Operations Agency (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 of Frontier Models 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 Attorney General 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.

---

[33] *Id.* at pp. 123-126.
[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.

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

> 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 and many aspects of that report have been integrated into this bill. Specifically, the Working Group report focused on transparency, incident reporting, and scoping. This bill incorporates major provisions and recommendations made by the Working Group to create a more narrowly focused framework to ensure foundation model safety and compliance.

Scoping. A major question that must be addressed before implementing any transparency measures 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 100 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, scheme, or course of conduct involving a "dangerous capability." This high threshold creates a regime in which only the most dire situations are captured. Furthermore, the bill categorizes dangerous capabilities into four distinct groups:

1. Provide expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon.
2. Conduct or assist in a cyberattack.
3. Engage in conduct, with limited human intervention, that would, if committed by a human, constitute the crime of extortion, theft, including theft by deception, or a serious or violent felony.
4. Evade 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 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.

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 alternatives 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, each of which emphasizes flexibility in how policy scope is determined. This bill requires large developers, defined as those with access to 10^{26} floating-point operations per second (FLOPS), a measure of computing power, to comply with the bill's transparency provisions. Notably, this threshold is tenfold higher than that set by the EU AI Act and matches one of the thresholds used in SB 1047. The bill applies to foundation models developed and deployed by these large developers. Foundation models, as defined in this bill, are AI systems trained on broad datasets, designed for general-purpose outputs, and adaptable to a variety of distinct tasks.

To ensure continual updating of this definition, the bill builds in flexibility regarding who qualifies as a large developer. Unlike with SB 1047, which established the Frontier Model Board, this bill vests the power to adjust the scope via regulation in the Attorney General (AG). Beginning in 2027, the AG is required to update the definition of a large developer to reflect technological advancements. As model training becomes more efficient, the amount of compute needed to create a model capable of catastrophic harm may decline. Furthermore, as the Working Group report notes, compute may not remain the most appropriate proxy for catastrophic risk. This bill allows the AG to adjust the definition of large developer to include more than just quantitative metrics, potentially incorporating thresholds based on model capabilities or monetary resources rather than raw compute alone.

In revising the definition, the AG 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 the resulting definition provides clarity for developers regarding their obligations under the law. Importantly, the bill specifies that if the AG determines that less-resourced developers, or those not currently operating at the frontier of AI development, nonetheless pose a substantial catastrophic risk, the AG must report this finding to the Legislature and offer a proposal for managing that risk. However, the AG may not include such developers within the definition of a large developer without further legislative action

---

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

> 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 of safety and security protocols (SSP), which large developers must draft, implement, and publish on their websites. The SSP must include:

- Criteria for Model Exclusion: A description of the criteria used to exclude certain foundation models from the SSP, along with the rationale for why those models do not pose a material threat.
- Risk Assessment Procedures: An explanation of how the developer assesses catastrophic risks, including those arising 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 SSP serves as a core transparency mechanism, ensuring that large developers maintain a baseline standard of safety practices to protect the public.

In tandem with the SSP, the bill also requires large 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

---

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

> 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 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.

This bill requires large 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 SSP. If a third party was engaged to assess the model, developers must disclose the same information, as well as details regarding the third party's predeployment access, independence, and any restrictions imposed by the developer on the assessment or on the third party's ability to disclose information.

Additionally, if the foundation model could pose a catastrophic risk, the transparency report must explain the rationale behind the decision to deploy the model, how that decision was made, and any limitations in the assessments used. Together, these requirements create a transparency report that lawmakers and outside experts can use to identify imminent risks and evaluate whether developers are employing appropriate assessments and interventions.

The bill also mandates that large developers publish on their websites any assessments of catastrophic risks arising from internal use of their foundation models, within the timeframe specified in their safety and security protocols. 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, this additional disclosure enables decisionmakers, experts, and the public to better understand and prepare for potential risks. 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 holistic transparency framework that will give insight and scrutiny to the processes from initial train 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:

---

[37] *Id.* at p. 19.
[38] *Id.* at p. 21.

> 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 bill strongly incorporates this recommendation by tasking the Attorney General (AG) with creating a mechanism for critical incident reporting. The bill defines a "critical incident" as any of the following:

- Unauthorized access to, modification of, or exfiltration of the model weights of a foundation model.
- Harm resulting from the materialization of a catastrophic risk.
- Loss of control of a foundation model causing death, bodily injury, or damage to rights in money or property.
- A foundation model using deceptive techniques against the large developer to subvert its controls or monitoring, outside the context of an evaluation designed to elicit such behavior.

Under this mechanism, a large developer or a member of the public may report a critical safety incident to the AG. Reports must include the date of the event, an explanation of how it qualifies as a critical incident, and a detailed description of the event. Large developers are required to report any critical safety incident within 15 days of discovering it. The AG must review all reports submitted by large 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.

CalCompute. This bill, similarly to 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

---

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

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 recent author's amendments expand the whistleblower protections to include disclosures concerning falsehoods or misrepresentations in the large developers' SSP or transparency reports.

Enforcement. The bill in print gives enforcement authority over this bill to the AG who can levy civil penalties against any large developer who violates this bill, for example, through making materially false or misleading statements about catastrophic risk from its foundation models or its management of catastrophic risk. The amount for the penalties have not yet been specified in the bill in print, but are fleshed out in the Committee amendments below.

---

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

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 SSP 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.

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 AG the authority to modify the definition of a "large developer" through rulemaking, a flexibility the RAISE Act does not include. While liability under SB 53 is currently unspecified, 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.

Arguably the most significant difference is the inclusion of third-party audits in the RAISE Act. Specifically, it requires large developers to undergo independent audits to assess compliance with its provisions, a requirement that was also included in the vetoed SB 1047. This provision reflects a key element of the Working Group Report that is not yet addressed in this bill. The passage of the RAISE Act in New York and its inclusion in the Working Group Report both underscore the potential importance of auditing in the broader artificial intelligence regulatory framework.

8) **Committee Amendments.** The author has agreed to various amendments that tighten and strengthen its provisions to increase accountability and enhance public trust.

Transparency. The transparency elements of this bill focus primarily on the steps that large developers will take to test for risks and the mitigations they will use to prevent such risks. However, this framing may be insufficient to ensure full transparency into the capabilities of these systems. For example, the Working Group Report highlights that Google Gemini's 2.5 Pro Model Card stated:

> The model's performance is strong enough that it has passed our early warning alert threshold, that is, we find it possible that subsequent revisions in the next few months could lead to a model that reaches the [critical capability level]"—with the "critical capability level" defined as a model that "can be used to significantly assist with high impact cyber attacks, resulting in overall cost/resource reductions of an order of magnitude or more"[42]

Similarly, the Working Group Report quotes OpenAI's April 2025 o3 and o4-mini System Card, which states:

---

[41] RAISE Act can be found at https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/A.
[42] Bommasani, *supra.* p.12.

> As we wrote in our deep research system card, several of our biology evaluations indicate our models are on the cusp of being able to meaningfully help novices create known biological threats, which would cross our high risk threshold. We expect current trends of rapidly increasing capability to continue, and for models to cross this threshold in the near future.[43]

In both instances, the developers focused on the inclusion of thresholds rather than on testing procedures to determine how their models perform or whether they pose a significant risk. These thresholds are ubiquitous throughout the foundation model ecosystem and provide insight into how developers gauge risk, which risks are prioritized in their workflows, and whether they are operating within the best practices of the field. Thresholds serve as benchmarks by which developers can identify when their models have achieved capabilities that require mitigation. Furthermore, capability thresholds add an additional layer of accountability, as outside experts can test models against those same thresholds to verify the accuracy of the developers' reports. This level of reproducibility engenders public trust.

Recognizing this, the author has agreed to amend the bill to require large developers to report the thresholds for both catastrophic risks and dangerous capabilities used in their SSP. Any threshold tests conducted, and the results thereof, must be documented in the transparency reports released at the time of deployment. Developers will also be required to disclose both the mitigations applied to those thresholds and the results of those mitigations. In addition, the author has agreed to amendments clarifying that transparency reports must include the reasoning behind the conclusions developers reach regarding risk or capability threshold assessments.

Lastly, the author has agreed to amend the bill to require disclosure, within the SSP, of whether the large developer has the capability to perform a full shutdown of the model in the event of a critical incident. **Recognizing that open-source or open-weight models are not under the control of the developer and thus cannot be fully shut down under any circumstance, this disclosure requirement applies only to developers who retain control over the model.** Unlike SB 1047, however, **this bill does not mandate the ability to perform a full shutdown**; rather, developers need only to disclose whether such a capability exists.

Scoping. This bill uses FLOPs as a metric to identify large developers—those with access to the most significant levels of compute. However, the Working Group Report notes that compute power alone is an imperfect basis for determining the appropriate scope of frontier model governance. It states:

> On this basis, we conclude that if training compute thresholds are used at all, they may function best when used as an initial filter to cheaply screen for entities that may warrant greater scrutiny. More broadly, within the family of cost-level metrics, training compute may still be the most attractive option, given other options are difficult to calibrate across different developers (e.g. dataset size, monetary cost), but they should actively monitor the rapid changes in how compute is allocated toward foundation model development and deployment.[44]

In response to these concerns, the author has agreed to amendments that add a second threshold to better define the intended scope. Under the amendments, a "large developer" must also have

---

[43] *Id.* at. p. 12.
[44] *Id.* at. p. 12.

annual gross revenues exceeding $100,000,000, a threshold recently referenced by Anthropic in its description of their transparency framework.[45] This change ensures that the bill's requirements apply only to major developers and do not place undue burdens on smaller startups.

Adverse Event Reporting. As stated above, under SB 53 critical incidents must be reported to the AG within 15 days. By contrast, the RAISE Act imposes a stricter 72-hour reporting requirement. This disparity raises important questions: Is prompt reporting necessary to ensure an effective response to a critical incident? And is the AG the appropriate authority to respond quickly to an imminent threat?

To address these concerns, the author has agreed to amend the bill to include a tiered system of reporting. Under this system, developers would be required to 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 the AG may not be. After alerting law enforcement, large developers would then still have 15 days to report the critical incident to the AG.

Furthermore, with the inclusion of capability thresholds in both the System Safety Plan (SSP) and transparency reports, the author has agreed to an additional amendment: if a model crosses certain capability thresholds designated by the large developer for the first time, this must also be reported to the AG.

A major goal of this bill is to increase transparency regarding the training and capabilities of foundation models. However, under the current framework, only the AG would have access to critical incident reports. This asymmetry in knowledge could lead to significant risks being overlooked by decisionmakers who are otherwise in a position to respond to such threats. As stated in the Working Group Report:

> In other domains, post-deployment monitoring by the government yields reports of medical complications, equipment malfunctions, near misses, and other unforeseen hazards. By aggregating information about incidents from developers and users, an adverse event reporting system that allows developers and downstream users to report post-deployment incidents promotes continuous learning, reduces information asymmetries between government and industry, and enables both regulators and industry to coordinate on mitigation efforts. As a result, an adverse event reporting system constitutes a critical first step toward data-driven assessments of the benefits and costs of regulatory actions.[46]

In response, the author has agreed to amend the bill to require the AG 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 large developer the report concerns.** Additionally, the amendments grant the AG 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.

---

[45] "The need for transparency in Frontier Ai", *Anthropic* (July 7, 2025), https://www.anthropic.com/news/the-need-for-transparency-in-frontier-ai.
[46] Bommasani, *supra.* p. 31.

Auditing. As noted above, the Working Group Report advocates for the use of independent third-party evaluations. Transparency elements require a "trust but verify" model, and by drawing lessons from historical precedents such as the energy sector the Working group states:

> The energy industry's internal documentation of speculative yet potentially irreversible consequences—illustrated by simulations predicting multiple future harms—does more than underscore the need for transparency. This case also reveals an added value of third-party oversight: The sheer magnitude of the social problem naturally invites further examination by governments and independent bodies. By aligning incentives and providing robust, independently validated evidence, enhanced third-party assessment could have empowered policymakers to make more thoughtful decisions aimed at mitigating long-term economic and societal damage.[47]

The report goes on to argue:

> Greater transparency is integral to an improved AI governance ecosystem; however, despite its considerable value, it is not sufficient without a broader ecosystem to complement it. Fundamentally, this is because the value of transparency from model developers is naturally limited by the information these developers have. For a nascent and complex technology being developed and adopted at a remarkably swift pace, developers alone are simply inadequate at fully understanding the technology and, especially, its risks and harms.
>
> Third-party risk assessment, therefore, is essential for building a more complete evidence base on the risks of foundation models and creating incentives for developers to increase the safety of their models. At present, some prominent foundation models are subject to risk evaluations by first-party model developers or second-party contractors. In contrast, third-party evaluation affords three distinctive strengths. First, third-party evaluations have unmatched scale: Thousands of individuals are willing to engage in risk evaluation, dwarfing the scale of internal or contracted teams. Second, third-party evaluations have unmatched diversity, especially when developers primarily reflect certain demographics and geographies that are often very different from those most adversely impacted by AI. Broad demographic, institutional, and disciplinary diversity is vital for unearthing blind spots. And finally, third-party evaluation is distinctively independent: Society requires forthright and trustworthy assessments of risk, which benefits from a lack of commercial and contractual entanglement with AI developers.
>
> A well-designed disclosure and verification regime offers significant benefits not only to the public but also to foundation model developers. By establishing industry-wide transparency standards and third-party verification mechanisms, companies can demonstrate compliance with best practices, potentially reducing their liability exposure compared to the uncertainties of purely reactive litigation. In practice, states could leverage the existing technical infrastructure and expertise of federal-level AI institutions to conduct standardized risk evaluations.
>
> Furthermore, this approach can transform competitive dynamics around safety. When safety measures and risk assessments are publicly disclosed and verified, companies face stronger

---

[47] *Id.* at p. 15.

> incentives to implement best practices, as deviations would be apparent and could attract greater scrutiny. This transparency coupled with third-party verification effectively creates a "race to the top" rather than a "race to the bottom" in safety practices, benefiting responsible companies while improving overall industry standards.[48]

Transparency is only meaningful when accompanied by a clear standard of accountability. Just as a teacher would not allow students to grade their own exams and expect complete honesty, the public should not expect every large developers to fully self-assess without independent oversight. The core objectives of this bill are to ensure the safe development and deployment of foundation models and to ensure that this is done in a responsible and verifiable manner. Achieving these goals will necessarily require third-party verification.

Nevertheless, concerns have been raised regarding the underdevelopment of the AI auditing market. During the March 2025 public comment period on the Working Group Draft, stakeholders noted the nascency of AI auditing as a field and expressed uncertainty about the current availability and scalability of qualified third-party auditors:

> Specific attention was also focused on third-party evaluations (sometimes referred to as audits in the feedback) by highlighting the immature third-party evaluation ecosystem, the lack of evaluation standards and broader measurement science, and the current barriers for evaluators to acquire access to the AI models and systems they evaluate.[49]

This concern has also been raised in the context of other measures proposing AI auditing requirements or establishing third-party auditing frameworks, such as AB 1018 (Bauer-Kahan), SB 420 (Padilla), and AB 1405 (Bauer-Kahan). However, this raises a broader policy question: **When is it appropriate to require third-party audits of large developers' practices?** As noted by the Working Group, in the energy sector, independent audits could have altered the trajectory of the climate crisis had they been implemented earlier. Reactive policies benefit no one except bad actors. Accountability in the development of foundation models should not hinge on a catastrophic safety incident. Instead, it should be a foundational element of any effective risk mitigation strategy.

Moreover, a core function of government, is to craft sound industrial policy that encourages the growth of new markets serving the public good when economic forces do not support that market. In this case, fostering a robust AI auditing ecosystem capable of verifying compliance with developers' SSPs and evaluating whether appropriate methodologies are being used to mitigate catastrophic risks is in the public interest.

Accordingly, the author has agreed to amend the bill to require large developers to submit their SSPs to annual audits conducted by independent third-party entities. To allow sufficient time for the AI auditing market to develop, this requirement will not take effect until 2030, providing a four-year runway for market readiness.

Large developers will be required to provide third-party auditors with all information necessary to evaluate both compliance with the SSP and whether the SSP is sufficiently clear to support

---

[48] *Id.* at p. 27-28.
[49] *Id.* at p. 42.

meaningful conclusions about the developer's safety practices. The final audit report must be retained by the developer, and a high-level summary of the audit must be submitted to the Attorney General within 30 days of completion. That summary will be included in the Attorney General's annual public report.

Ultimately, the inclusion of independent third-party audits is essential to advancing the bill's overarching objective: the creation of a comprehensive and accountable transparency regime.

Enforcement. As currently drafted, the bill includes vague provisions for civil penalties to be administered by the Attorney General (AG) for violations of the bill. To address this ambiguity, the author has agreed to amend the bill to establish a tiered system of enforcement.

Under the amended framework, the lowest tier of violations would apply to unknowing violations of the bill that do not create a danger of death, serious physical injury, or a catastrophic risk—for example, the late publication of a required report. These violations would be subject to a civil penalty but would not trigger immediate enforcement action. The next tier includes either knowing violations that do not result in a threat to life or safety or unknowing violations that do create a danger of death, serious physical injury, or a catastrophic risk. These violations also carry civil penalties but reflect a higher level of potential harm or culpability. The highest tier is reserved for knowing violations that do create a danger of death, serious physical injury, or a catastrophic risk. These represent the most serious offenses under the bill and would be subject to the most significant penalties. Penalties will scale according to the level of violation. For the lowest tier of offenses, the bill clarifies that developers must be provided a 30-day grace period to cure the violation after receiving notice from the AG before penalties may be imposed.

Finally, the amended bill extends enforcement provisions to include violations committed by auditors, who will also be subject to civil penalties for noncompliance.

Clarifying amendments. The author has also agreed to various clarifying, technical, and cleanup amendments. The amendments are shown below:

SECTION 1. The Legislature finds and declares all of the following:

(a) California is leading the world in artificial intelligence innovation and research through companies large and small and through the state's remarkable public and private universities.

(b) Artificial intelligence, including new advances in foundation models, has the potential to catalyze innovation and the rapid development of a wide range of benefits for Californians and the California economy, including advances in medicine, wildfire forecasting and prevention, and climate science, and to push the bounds of human creativity and capacity.

(c) The Joint California Policy Working Group on AI Frontier Models has recommended sound principles for policy in artificial intelligence.

(d) Targeted interventions to support effective artificial intelligence governance should balance the technology's benefits and material risks.

(e) Artificial intelligence developers have already voluntarily committed to creating safety and security protocols and releasing the results of risk assessments.

(f) In building a robust and transparent evidence environment, policymakers can align incentives to simultaneously protect consumers, leverage industry expertise, and recognize leading safety practices.

(g) When industry actors conduct internal research on their technologies' impacts, a significant information asymmetry can develop between those with privileged access to data and the broader public.

(h) Greater transparency, given current information deficits, can advance accountability, competition, and public trust.

(i) Whistleblower protections and public-facing information sharing are key instruments to increase transparency.

(j) Adverse event reporting systems enable monitoring of the post-deployment impacts of artificial intelligence.

(k) There is growing evidence that, unless they are developed with careful diligence and reasonable precaution, advanced artificial intelligence systems could pose catastrophic risks from both malicious uses and malfunctions, including artificial intelligence-enabled hacking, biological attacks, and loss of control.

(l) With the frontier of artificial intelligence rapidly evolving, there is a need for legislation to track the frontier of artificial intelligence research and alert policymakers and the public to the risks and harms from the very most advanced artificial intelligence systems, while avoiding burdening smaller companies behind the frontier.

~~**(m) A computational threshold of 10^26 floating point operations captures the current frontier of foundation model development and captures only highly resourced developers spending hundreds of millions of dollars to develop foundation models.**~~

(~~**n**~~***m***) In the future, foundation models developed by smaller companies or that are behind the frontier may pose significant catastrophic risk, and additional legislation may be needed at that time.

(~~**o**~~***n***) It is the intent of the Legislature to create more transparency, but collective safety will depend in part on large developers taking due care in their development and deployment of foundation models proportional to the scale of the foreseeable risks.

SEC. 2. Chapter 25.1 (commencing with Section 22757.10) is added to Division 8 of the Business and Professions Code, to read:

CHAPTER 25.1. Transparency in Frontier Artificial Intelligence Act

22757.10. This chapter shall be known as the Transparency in Frontier Artificial Intelligence Act.

22757.11. For purposes of this chapter:

(a) "Artificial intelligence model" means 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" means a foreseeable and material risk that a large developer's development, storage, use, or deployment of a foundation model will materially contribute to the death of, or serious injury to, more than ~~100~~ ***50*** people or more than one billion dollars ($1,000,000,000) in damage to ~~rights in money or~~ ***,or loss of,*** property arising from a single incident, scheme, or course of conduct involving a dangerous capability.

(c) "Critical safety incident" means any of the following:

(1) Unauthorized access to, modification of, or exfiltration of, the model weights of a foundation model.

(2) Harm resulting from the materialization of a catastrophic risk.

(3) Loss of control of a foundation model causing death, bodily injury, or damage to ~~rights in money,~~ or ***loss of,*** property.

(4) A foundation model that uses deceptive techniques against the large developer to subvert the controls or monitoring of its large developer outside of the context of an evaluation designed to elicit this behavior.

(5) Attaining a dangerous capability or catastrophic risk threshold, as defined in the large developer's safety and security protocol pursuant to paragraph (3) of subdivision (a) of Section 22757.12, for the first time.

(d) "Dangerous capability" means the capacity of a foundation model to do any of the following:

(1) Provide expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon.

(2) Conduct or assist in a cyberattack.

(3) Engage in conduct, with limited human intervention, that would, if committed by a human, constitute the crime of ***murder, assault,*** extortion, theft, including theft by ***false pretense*** ~~deception, or a serious or violent felony~~.

(4) Evade the control of its large developer or user.

(e) (1) "Deploy" means to make a foundation model available to a third party for use, modification, copying, or combination with other software.

(2) "Deploy" does not include making a foundation model available to a third party for the primary purpose of developing or evaluating the foundation model.

(f) "Large developer" means either of the following:

(1) (A) Before January 1, 2027, "large developer" means a person who ~~**has trained, or initiated the training of, at least one foundation model using a quantity of computing power greater than 10^26 integer or floating-point operations**~~. ***meets both of the following criteria:***

(i) (I) The person has trained, or initiated the training of, at least one foundation model using a quantity of computing power greater than 10^26 integer or floating-point operations.

(II) The quantity of computing power described in ~~subparagraph (A)~~ ***subclause (I)*** shall include computing for the original training run and any subsequent fine-tuning, reinforcement learning, or other material modifications to a preceding foundation model.

(ii) The person had annual gross revenues in excess of one hundred million dollars ($100,000,000) in the preceding calendar year.

(B) If the Attorney General does not adopt a regulation described in subparagraph (A) by January 1, 2027, the definition in paragraph (1) shall be operative until the regulation is adopted.

(g) "Foundation model" means an artificial intelligence model that is all of the following:

(1) Trained on a broad data set.

(2) Designed for generality of output.

(3) Adaptable to a wide range of distinctive tasks.

(h) "Model weight" means a numerical parameter in a foundation model that is adjusted through training and that helps determine how inputs are transformed into outputs.

(i) "Property" means tangible or intangible property.

(~~i~~***j***) "Safety and security protocol" means documented technical and organizational protocols to manage, assess, and mitigate catastrophic risks.

22757.12. (a) A large developer shall write, implement, and clearly and conspicuously publish on its internet website a safety and security protocol that describes in specific detail all of the following:

(1) How, if at all, the large developer excludes certain foundation models from being covered by its safety and security protocol because those foundation models ***are incapable of*** ~~**do not**~~ **posinge** material catastrophic risks ***and which foundation models, if any, are included***.

(2) The testing procedures that the large developer uses to assess catastrophic risks from its foundation models, including risk resulting from malfunctions, misuse, and foundation models evading the control of the large developer or user.

(3)***(A) Thresholds used by the large developer to identify and assess whether the foundation model has attained a dangerous capability or poses a catastrophic risk.***

(B) How the large developer will assess whether those thresholds have been attained, which may include multiple tiered thresholds for different dangerous capabilities or for catastrophic risk.

(C)The actions the large developer will take if each threshold is attained.

(~~3~~***4***) The mitigations that a large developer takes to reduce a catastrophic risk and how the large developer assesses the effectiveness of those mitigations.

(~~4~~***5***) The degree to which the large developer's assessments of catastrophic risk ***and dangerous capabilities*** and the effectiveness of catastrophic risk mitigations are reproducible by external entities.

(~~5~~6) The extent to which, and how, a large developer will use third parties to assess catastrophic risks ***and dangerous capabilities*** and the effectiveness of mitigations of catastrophic risk.

(~~6~~***7***) The large developer's cybersecurity practices and how the large developer secures unreleased model weights from unauthorized modification or transfer by internal or external parties.

(~~7~~***8***) To the extent that the foundation model is controlled by the large developer, the procedures the large developer will use to monitor critical safety incidents and the steps that a large developer would take to respond to a critical safety incident, including, but not limited to, ***whether the large developer has the ability to promptly shut down copies of foundation models owned and controlled by the large developer,*** who the large developer will notify and the timeline on which the large developer would take these steps.

(~~8~~***9***) The testing procedures that the large developer will use to assess and manage a catastrophic risk ***or dangerous capability*** resulting from the internal use of its foundation models, including risks resulting from a foundation model circumventing oversight mechanisms, and the schedule, specified in days, by which the large developer will report these assessments pursuant to subdivision (d).

(~~9~~***10***) How the developer determines when its foundation models are substantially modified enough to conduct additional assessments and publish a transparency report pursuant to subdivision (c).

(b) If a large developer makes a material modification to its safety and security protocol, the large developer shall clearly and conspicuously publish the modified protocol and a justification for that modification within 30 days.

(c) Before or concurrently with deploying a new foundation model or a substantially modified version of an existing foundation model ***covered by the large developer's safety and security protocol***, a large developer shall clearly and conspicuously publish on its internet website a transparency report containing all of the following:

(1)***(A)*** The results of any risk assessment ***or risk mitigation assessment conducted by the large developer or a third party that contracts with a large developer pursuant to its safety and security protocol, why the information gathered by the large developer or third party leads to the stated results, and*** the steps taken to address any identified risks~~**, and the results of any risk mitigation assessment conducted by the large developer pursuant to its safety and security protocol**~~.

~~**(2) (A) The results of any risk assessment, the steps taken to address any identified risks, and the results of any risk mitigation assessment conducted pursuant to a large developer's safety and security protocol that is conducted by a third party that contracts with a large developer.**~~

(B) A large developer shall disclose the time and extent of predeployment access provided to any third party described in subparagraph (A), whether or not the third party was independent, and the nature of any constraints the large developer placed on the assessment or on the third party's ability to disclose information about its assessment to the public or to government officials.

(C) Whether a catastrophic risk threshold or dangerous capability threshold has been attained for the foundation model and any actions taken as a result.

(~~3~~***2***) (A) ~~**If the deployment would pose a catastrophic risk, t**~~***T***he reasoning behind the large developer's decision to deploy the foundation model, the process by which the large developer arrived at that decision, and any limitations in the assessments that the large developer used to make that decision.

(B) A large developer may reuse an answer previously provided under subparagraph (A) if the rationale in question has not materially changed for the new deployment.

(d) A large developer shall clearly and conspicuously publish on its internet website any assessment of catastrophic risk ***or dangerous capabilities*** resulting from internal use of its foundation models pursuant to the schedule the developer specifies in its safety and security protocol.

(e) A large developer shall not make a materially false or misleading statement about catastrophic risk from its foundation models**,** ~~**or**~~ its management of catastrophic risk***, or its implementation of or compliance with its safety and security protocol***.

(f) (1) When a large developer publishes documents to comply with this section, the large developer may make redactions to those documents that are necessary to protect the large developer's trade secrets, the large developer's cybersecurity, public safety, or the national security of the United States or to comply with any federal or state law.

(2) If a large developer redacts information in a document pursuant to this subdivision, the large developer shall 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.

22757.13. (a) The Attorney General shall establish a mechanism to be used by a large developer or a member of the public to report a critical safety incident that includes all of the following:

(1) The date of the critical safety incident.

(2) The reasons the incident qualifies as a critical safety incident.

(3) A short and plain statement describing the critical safety incident.

(b) A large developer shall report any critical safety incident pertaining to one or more of its foundation models to the Attorney General within 15 days of discovering the critical safety incident.

(c) If a large developer discovers a critical safety incident poses an imminent risk of danger of death or serious physical injury, the large developer shall disclose that incident as soon as practicable, and in no event later than 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 in accordance with applicable law.

(~~c~~***d***) The Attorney General shall review critical safety incident reports submitted by large developers and may review reports submitted by members of the public.

(e) The Attorney General may transmit reports of critical safety incidents, reports from employees pursuant to Chapter 5.1 (commencing with Section 1107) of Part 3 of Division 2 of the Labor Code, and summaries of auditor's reports pursuant to Section 22757.16 to the Legislature, Governor, the federal government or appropriate state agencies at the Attorney General's discretion and in accordance with relevant law.

(f) A report of a critical safety incident submitted to the Attorney General pursuant to this section, reports from employees pursuant to Chapter 5.1 (commencing with Section 1107) of Part 3 of Division 2 of the Labor Code, and summaries of auditor's reports pursuant to Section 22757.16 are exempt from the California Public Records Act (Division 10 (commencing with Section 7920.000) of Title 1 of the Government Code).

(g) (1) Beginning January 1, 2027, and annually thereafter, the Attorney General shall produce a report with anonymized and aggregated information about critical safety incidents, reports from employees pursuant to Chapter 5.1 (commencing with Section 1107) of Part 3 of Division 2 of the Labor Code, and summaries of auditor's reports pursuant to Section 22757.16 that have been reviewed by the Attorney General since the preceding report.

(2) The Attorney General shall not include information in a report pursuant to this subdivision that would compromise the trade secrets or cybersecurity of a large developer, public safety, or the national security of the United States or that would be prohibited by any federal or state law.

(3) The Attorney General shall transmit a report pursuant to this subdivision to the legislature pursuant to Section 9795 and to the Governor.

22757.14. ~~**(a) A violation of this chapter shall be subject to a civil penalty in an amount not to exceed ____ dollars ($____) per violation, unless the violation is willful or reckless, in which case the civil penalty shall be in an amount not to exceed ____ dollars ($____).**~~

(a) A violation of this chapter by a large developer shall be subject to a civil penalty as follows:

(1) (A) For an unknowing violation that does not create a material risk of death, serious physical injury, or a catastrophic risk, a civil penalty in an amount not to exceed ten thousand dollars ($10,000).

(B) (i) In the case of a first violation subject to a civil penalty pursuant to subparagraph (A), a large developer shall be provided with a thirty (30) day right to cure the violation after notification by the Attorney General. If the violation is cured within that period, no civil penalty shall be imposed for that violation.

(ii) For the purposes of this subparagraph, a violation involving the missing or late publication or submission of a document is deemed cured when the developer publishes or submits that document, even if the publication or submission occurs after the statutory deadline.

(2) For a knowing violation that does not create a material risk of death, serious physical injury, or a catastrophic risk or an unknowing violation that creates a material risk of death, serious physical injury, or a catastrophic risk, a civil penalty in an amount not to exceed one hundred thousand dollars ($100,000).

(3) For a knowing violation that creates a material risk of death, serious physical injury, or a catastrophic risk, a civil penalty in an amount not to exceed one million dollars ($1,000,000) for a first such violation and in an amount not exceeding ten million dollars ($10,000,000) for any subsequent such violation.

(b) A violation of this chapter by an auditor shall be subject to a civil penalty in an amount not to exceed ten thousand dollars ($10,000).

(~~b~~***c***) The civil penalty shall be assessed in a civil action brought only by the Attorney General.

22757.15. (a) On or before January 1, 2027, and annually thereafter, the Attorney General may adopt regulations to update the definition of a "large developer" for the purposes of this chapter to ensure that it accurately reflects technological developments, scientific literature, and widely accepted national and international standards and applies to well-resourced large developers at the frontier of artificial intelligence development.

(b) In developing regulations pursuant to this section, the Attorney General shall take into account all of the following:

(1) Similar thresholds used in international standards or federal law, guidance, or regulations for the management of catastrophic risk.

(2) Input from stakeholders, including academics, industry, the open-source community, and governmental entities.

(3) The extent to which a person will be able to determine, before beginning to train or deploy a foundation model, whether that person will be subject to the regulations as a large developer with an aim toward allowing earlier determinations if possible.

(4) The complexity of determining whether a person is covered, with an aim toward allowing simpler determinations if possible.

(5) The external verifiability of determining whether a person is covered, with an aim toward definitions that are verifiable by parties other than the large developer.

(c) If the Attorney General determines that less well-resourced developers, or developers significantly behind the frontier of artificial intelligence, may create substantial catastrophic risk, the Attorney General shall promptly submit a report to the Legislature, pursuant to Section 9795, with a proposal for managing this source of catastrophic risk but shall not include those developers within the definition of "large developer" without authorization in subsequently enacted legislation.

22757.16. (a) Beginning January 1st, 2030, and at least annually thereafter, a large developer shall retain an independent third-party auditor to produce a report assessing both of the following:

(1) Whether the large developer has substantially complied with its safety and security protocol and any instances of substantial noncompliance during the prior year.

(2) Any instances where the large developer's safety and security protocol has not been stated clearly enough to determine whether the large developer has complied.

(b) A large developer shall allow the third-party auditor access to all materials produced to comply with this Act and any other materials reasonably necessary to perform the assessment required under subsection (a).

(c) A large developer shall retain the auditor's report for 5 years.

(d) In conducting the audit, the auditor shall employ or contract one or more individuals with expertise in corporate compliance and one or more individuals with technical expertise in the safety of foundation models.

(e) Within 30 days after completing an audit, the auditor shall transmit to the Attorney General a high-level summary of the auditor's report.

(f) The high-level summary required by subdivision (e) shall fairly present, in all material respects, the outcome of the audit, and an auditor shall not knowingly include a material misrepresentation or omit a material fact in either that summary or the auditor's report.

SEC. 3. Section 11546.8 is added to the Government Code, to read:

11546.8. (a) There is hereby established within the Government Operations Agency a consortium that shall develop, pursuant to this section, a framework for the creation of a public cloud computing cluster to be known as "CalCompute."

(b) The consortium shall develop a framework for the creation of 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:

(1) Fostering research and innovation that benefits the public.

(2) Enabling equitable innovation by expanding access to computational resources.

(c) The consortium shall make reasonable efforts to ensure that CalCompute is established within the University of California to the extent possible.

(d) CalCompute shall include, but not be limited to, all of the following:

(1) A fully owned and hosted cloud platform.

(2) Necessary human expertise to operate and maintain the platform.

(3) Necessary human expertise to support, train, and facilitate the use of CalCompute.

(e) The consortium shall operate in accordance with all relevant labor and workforce laws and standards.

(f) (1) On or before January 1, 2027, the Government Operations Agency shall submit, pursuant to Section 9795, a report from the consortium to the Legislature with the framework developed pursuant to subdivision (b) for the creation and operation of CalCompute.

(2) The report required by this subdivision shall include all of the following elements:

(A) A landscape analysis of California's current public, private, and nonprofit cloud computing platform infrastructure.

(B) An analysis of the cost to the state to build and maintain CalCompute and recommendations for potential funding sources.

(C) Recommendations for the governance structure and ongoing operation of CalCompute.

(D) Recommendations for the parameters for use of CalCompute, including, but not limited to, a process for determining which users and projects will be supported by CalCompute.

(E) An analysis of the state's technology workforce and recommendations for equitable pathways to strengthen the workforce, including the role of CalCompute

(F) A detailed description of any proposed partnerships, contracts, or licensing agreements with nongovernmental entities, including, but not limited to, technology-based companies, that demonstrates compliance with the requirements of subdivisions (c) and (d).

(G) Recommendations regarding how the creation and ongoing management of CalCompute can prioritize the use of the current public sector workforce.

(g) The consortium shall, consistent with state constitutional law, consist of 14 members as follows:

(1) 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.

(2) Three representatives of impacted workforce labor organizations appointed by the Speaker of the Assembly.

(3) 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.

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

(h) The members of the consortium shall serve without compensation, but shall be reimbursed for all necessary expenses actually incurred in the performance of their duties.

(i) The consortium shall be dissolved upon submission of the report required by paragraph (1) of subdivision (f) to the Legislature.

(j) If CalCompute is established within the University of California, the University of California may receive private donations for the purposes of implementing CalCompute.

(k) This section shall become operative only upon an appropriation in a budget act, or other measure, for the purposes of this section.

SEC. 4. Chapter 5.1 (commencing with Section 1107) is added to Part 3 of Division 2 of the Labor Code, to read:

CHAPTER 5.1. Whistleblower Protections: Catastrophic Risks in AI Foundation Models

1107. For purposes of this chapter:

(a) "Artificial intelligence model" means 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" has the meaning defined in Section 22757.11 of the Business and Professions Code.

(c) "Large developer" has the meaning defined in Section 22757.11 of the Business and Professions Code.

(d) "Employee" means a person who performs services for an employer, including both of the following:

(1) A contractor, subcontractor, or an unpaid advisor involved with assessing, managing, or addressing catastrophic risk, including all of the following:

(A) An independent contractor.

(B) A freelance worker.

(C) A person employed by a labor contractor.

(D) A board member.

(2) Corporate officers.

(e) "Foundation model" has the same meaning as defined in Section 22757.11 of the Business and Professions Code.

(f) "Labor contractor" means an individual or entity that supplies, either with or without a contract, a client employer with workers to perform labor within the client employer's usual course of business.

1107.1. (a) A large developer shall not make, adopt, enforce, or enter into a rule, regulation, policy, or contract that prevents an employee from disclosing, or retaliates against an employee for disclosing, information to the Attorney General, a federal authority, a person with authority over the employee, or another employee who has authority to investigate, discover, or correct the reported issue, if the employee has reasonable cause to believe that the information discloses either of the following:

(1) The large developer's activities pose a catastrophic risk.

(2) The large developer has violated Chapter 25.1 (commencing with Section 22757.10) of Division 8 of the Business and Professions Code.

(b) A large developer shall not enter into a contract that prevents an employee from making a disclosure protected under Section 1102.5.

(c) A large developer shall not make, adopt, enforce, or enter into a rule, regulation, policy, or contract that would prevent an organization or entity that provides goods or services to the large developer related to the assessment, management, or addressing of catastrophic risk, or an employee of that organization or entity, from disclosing information to the Attorney General, a federal authority, or the developer if the organization, entity, or individual has reasonable cause to believe that the information discloses either of the following

(1) The large developer's activities pose a catastrophic risk.

(2) The large developer has violated Chapter 25.1 (commencing with Section 22757.10) of Division 8 of the Business and Professions Code.

(d) An employee may use the hotline described in Section 1102.7 to make reports described in subdivision (a).

(e) A large developer shall provide a clear notice to all employees of their rights and responsibilities under this section, including by doing either of the following:

(1) At all times posting and displaying within any workplace maintained by the large developer a notice to all employees of their rights under this section, ensuring that any new employee receives equivalent notice, and ensuring that any employee who works remotely periodically receives an equivalent notice.

(2) At least once each year, providing written notice to each employee of the employee's rights under this section and ensuring that the notice is received and acknowledged by all of those employees.

(f) (1) A large developer shall provide a reasonable internal process through which an employee may anonymously disclose information to the large developer if the employee believes in good faith that the information indicates that the large developer's activities present a catastrophic risk or that the large developer violated Chapter 25.1 (commencing with Section 22757.10) of

Division 8 of the Business and Professions Code, including a monthly update to the person who made the disclosure regarding the status of the large developer's investigation of the disclosure and the actions taken by the large developer in response to the disclosure.

(2) (A) Except as provided in subparagraph (B), the disclosures and responses of the process required by this subdivision shall be shared with officers and directors of the large developer at least once each quarter.

(B) If an employee has alleged wrongdoing by an officer or director of the large developer in a disclosure or response, subparagraph (A) shall not apply with respect to that officer or director.

(g) The court is authorized to award reasonable attorney's fees to a plaintiff who brings a successful action for a violation of this section.

(h) In a civil action brought pursuant to this section, once it has been demonstrated by a preponderance of the evidence that an activity proscribed by this section was a contributing factor in the alleged prohibited action against the employee, the large developer shall have the burden of proof to demonstrate by clear and convincing evidence that the alleged action would have occurred for legitimate, independent reasons even if the employee had not engaged in activities protected by this section.

(i) (1) In a civil action or administrative proceeding brought pursuant to this section, an employee may petition the superior court in any county wherein the violation in question is alleged to have occurred, or wherein the person resides or transacts business, for appropriate temporary or preliminary injunctive relief.

(2) Upon the filing of the petition for injunctive relief, the petitioner shall cause notice thereof to be served upon the person, and thereupon the court shall have jurisdiction to grant temporary injunctive relief as the court deems just and proper.

(3) In addition to any harm resulting directly from a violation of this section, the court shall consider the chilling effect on other employees asserting their rights under this section in determining whether temporary injunctive relief is just and proper.

(4) Appropriate injunctive relief shall be issued on a showing that reasonable cause exists to believe a violation has occurred.

(5) An order authorizing temporary injunctive relief shall remain in effect until an administrative or judicial determination or citation has been issued, or until the completion of a review pursuant to subdivision (b) of Section 98.74, whichever is longer, or at a certain time set by the court. Thereafter, a preliminary or permanent injunction may be issued if it is shown to be just and proper. Any temporary injunctive relief shall not prohibit a large developer from disciplining or terminating an employee for conduct that is unrelated to the claim of the retaliation.

(j) Notwithstanding Section 916 of the Code of Civil Procedure, injunctive relief granted pursuant to this section shall not be stayed pending appeal.

(k) (1) This section does not impair or limit the applicability of Section 1102.5.

(2) The remedies provided by this section are cumulative to each other and the remedies or penalties available under all other laws of this state.

SEC. 5. (a) The provisions of this act are severable. If any provision of this act or its application is held invalid, that invalidity shall not affect other provisions or applications that can be given effect without the invalid provision or application.

(b) This act shall be liberally construed to effectuate its purposes.

(c) The duties and obligations imposed by this act are cumulative with any other duties or obligations imposed under other law and shall not be construed to relieve any party from any duties or obligations imposed under other law and do not limit any rights or remedies under existing law.

(d) This act shall not apply to the extent that it strictly conflicts with the terms of a contract between a federal government entity and a large developer.

(e) This act shall not apply to the extent that it is preempted by federal law.

SEC. 6. The Legislature finds and declares that Section 2 of this act, which adds Chapter 25.1 (commencing with Section 22757.10) to Division 8 of the Business and Professions Code, imposes a limitation on the public's right of access to the meetings of public bodies or the writings of public officials and agencies within the meaning of Section 3 of Article I of the California Constitution. Pursuant to that constitutional provision, the Legislature makes the following findings to demonstrate the interest protected by this limitation and the need for protecting that interest:

Information in critical safety incident reports, summaries of auditor's reports, and reports from employees may contain information that could threaten public safety or compromise the response to an incident if disclosed to the public.

ARGUMENTS IN SUPPORT: Secure AI Project, co-sponsors of the bill, alongside a coation of technology equity advocacy groups write in support:

> The California Report on Frontier AI Policy, while it does not endorse any specific legislation, forms the foundation for SB 53. Established by Governor Newsom in 2024 and led by Dr. Fei-Fei Li, Dr. Jennifer Tour Chayes and Mariano-Florentino Cuéllar, the report is anchored on the notion of "trust but verify" and calls for more transparency into the safety practices of AI companies, adverse event reporting requirements, and whistleblower protections. SB 53 implements these principles.
>
> Large AI developers are developing increasingly advanced AI systems. We are excited about the potential for these systems to drive improvements in education, science, provisioning of public services, and more. At the same time, large AI developers themselves warn that their AI systems could pose serious risks, which they have voluntarily committed to addressing. The Report stated that "some risks have unclear but growing evidence...AI-enabled hacking or biological attacks, and loss of control" – the risks that SB 53 aims to address and gather more evidence about. Advanced AI is currently mostly unregulated, and these risks are currently being managed by companies themselves without any requirement that they inform

the public about their risk management practices or report serious incidents. SB 53 addresses this much needed gap by implementing four key recommendations from the report.

First, the Report argued that "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." SB 53 implements this recommendation as a requirement for large AI developers to write, publish, and follow safety and security protocols to manage the most severe risks. This is in line with voluntary commitments that companies have already made. Rather than prescribe specific technical standards that companies must take, the bill simply requires companies to be transparent about the approaches they are using. Some of the specific required elements of safety protocols, such as a requirement to manage risks related to internal use of AI models and cybersecurity policies, directly mirror recommendations in the Report. Others mirror components of the Stanford Foundation Model Transparency Index, which is cited prominently in the Report.

Second, the Report stated that "transparency into pre-deployment assessments of capabilities and risks, spanning both developer-conducted and externally conducted evaluations, is vital given that these evaluations are early indicators of how models may affect society and may be interpreted (potentially undesirably) as safety assurances." SB 53 accomplishes this with a requirement that large developers publish transparency reports that include the results of their pre-deployment assessments of catastrophic risk. The Report also argues that "transparency into the safety cases used to assess risk provides clarity into how developers justify decisions around model safety," which forms the basis for 22757.12(c)(3).

Third, the Report concluded that "an adverse event reporting system that combines mandatory developer reporting with voluntary user reporting maximally grows the evidence base." SB 53 takes exactly this approach by establishing a tightly defined set of critical safety incidents that AI developers are required to report to the Attorney General. It would also allow members of the public to optionally submit reports.

Finally, the Report recommends strengthening whistleblower protections, pointing out that "actions that may clearly pose a risk and violate company policies...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." This recommendation is mirrored in SB 53, which allows employees to report evidence of catastrophic risks as well as violations of SB 53 itself to government authorities with legal protections against retaliation.

SB 53 only applies to the largest AI developers – those training models with more than 10^26 floating point operations (FLOPs). These are companies spending hundreds of millions or billions of dollars to train the most advanced AI models. It would impose no burden on smaller companies and the requirements it imposes on large companies are minimal compared to what companies are already voluntarily doing. The Report argues that "policymakers should ensure that mechanisms are in place to adapt thresholds over time—not only by updating specific threshold values but also by revising or replacing metrics if needed." It also suggests specific criteria that thresholds should be evaluated for. Following this recommendation, SB 53 allows the Attorney General to update the definition of "large developer" through regulation while considering the same factors described in the report.

Regardless of any update, the Attorney General must only include "well-resourced large developers at the frontier of artificial intelligence development" in the scoping of the bill. If legislation is needed to cover other developers, the Attorney General is instructed to write a report to the Legislature requesting it.

Finally, SB 53 would also set in motion CalCompute, a public cloud computing cluster for use by academics and startups in California. Computational resources are essential for AI research and CalCompute would make those resources more accessible to California's top universities and startups, helping to catalyze additional research into beneficial applications of AI and supporting, in particular, smaller startups for a healthier innovation ecosystem. This mirrors a similar computing cluster that is already being established in New York state. We support this groundbreaking effort, which would advance and democratize AI research in California.

SB 53 thoughtfully implements the recommendations of the Report by combining a low-burden transparency and reporting regime with a public compute cluster that will broaden access for AI researchers and startups in California. This is a commonsense approach that will strengthen the AI ecosystem, benefiting both companies and the public interest.

For all these reasons, we respectfully urge your support of this important measure.

ARGUMENTS IN OPPOSITION: In opposition to the bill, Chamber of Progress argues:

On behalf of the Chamber of Progress, a tech industry association supporting public policies to build a more inclusive society in which all people benefit from technological advances, we respectfully urge you to oppose SB 53, based on its recent amendments.

The definition of "catastrophic risk" remains vague and overreaching

While the amended bill replaces the term "critical risk" with "catastrophic risk," the underlying problem persists. The definition remains overly expansive and ambiguous, capturing a wide array of hypothetical scenarios that may not reflect real-world AI capabilities or threats.

Under Section 22757.11(b), the definition of "catastrophic risk" includes scenarios where a foundation model is "materially likely" to cause harm, potentially due to misuse or malicious inputs. However, this standard is vague, lacks clear and objective thresholds, and leaves room for subjective interpretation by whistleblowers or regulators. In a rapidly evolving field like AI, such ambiguity could unfairly penalize developers who are acting responsibly.

In addition, the inclusion of highly abstract risks, such as the evasion of human control under Section 22757.12(a)(2), creates significant uncertainty. Without clear technical criteria, companies may face liability or investigation based on assumptions about what a model might enable rather than what it has demonstrably done. This uncertainty undermines research and commercial deployment in California and could push critical AI development efforts out of state or abroad.

The $100,000,000 compute cost threshold risks misidentifying frontier AI models

SB 53's use of an arbitrary $100,000,000 compute cost threshold to determine eligibility for protections is an inherently flawed method for identifying frontier AI models. This threshold may result in the overinclusion of developers working on benign systems while potentially excluding smaller models that pose significant real-world risks. It also ignores the constantly changing cost of compute.

A more effective approach would involve a threshold based on model capabilities, deployment context, and specific use cases rather than relying solely on computational costs.

SB 53's extensive safety and security protocols create impractical burdens for AI developers

SB 53 imposes comprehensive safety and security requirements on AI developers, as outlined in Section 22757.12(a), including risk testing, deployment practices, and escalation procedures. While these objectives are important, the bill demands an impractical level of detailed planning and documentation for every conceivable misuse scenario, many of which are speculative or unrealistic.

This exhaustive approach compels developers to allocate significant time and resources toward preparing for hypothetical risks rather than addressing actual, demonstrable harms. For startups and smaller companies, these extensive protocols create a heavy administrative burden that diverts critical resources away from innovation and the timely deployment of beneficial AI technologies.

Additionally, Section 22757.12(c)'s requirement that developers publish detailed transparency reports, before or at the time of deploying a new or substantially modified foundation model, creates significant risks to both competitiveness and operational security.

Although redactions are permitted under subsection (f), the requirement to publish the "character and justification" of redacted material could still inadvertently expose business-sensitive strategies or vulnerabilities. This level of forced transparency goes beyond reasonable accountability and may discourage responsible companies from operating in California. It also creates opportunities for misuse by malicious actors who could exploit disclosed model weaknesses or mitigation gaps.

In fast-moving AI markets, publication of this level of detail erodes a developer's ability to maintain a competitive edge and deters innovation by raising legal and reputational risks associated with even speculative harms.

Taking an opposed unless amended position, the California Chamber of Commerce in a coalition with other technology trade organizations argue:

As a general matter, we see areas where the bill diverges from the final findings of Governor Newsom's Joint California Policy Working Group on Frontier Models, which arose out of his veto of SB 1047 (Wiener, 2024) and have drawn your attention to certain areas where alignment can be better sought. While we cannot support **SB 53** in its current form, we anticipate that amendments can be made to address these issues. While these are only our initial views given the complexity of the issues raised by the amendments, we hope there is the opportunity to work on this legislation with you to get this important policy right. In the interim, we believe this provides a strong overview of a path forward from our perspective to

ensure **SB 53** will meaningfully advance AI safety; rather than risk imposing substantial compliance costs and technical and legal challenges for certain AI developers to assess and prevent any and all downstream risks once their models are available for others to modify and deploy.

SB 53 should ideally focus on model risk, not developer size—otherwise, it will likely fail to address concerns about smaller, specialized models being a source of catastrophic risk

We are concerned about the scope of the bill and its focus on "large developers" to the exclusion of other developers of models with advanced capabilities that pose risks of catastrophic harm. In his veto of SB 1047 last year, Governor Newsom criticized this approach stating:

> 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.

As amended July 8th, **SB 53** is neither sufficiently focused on, nor does it provide any definition of what constitutes a "frontier" model – effectively equating today's frontier AI models with foundation models and defining the threshold for covered models from there. Specifically, under the premise that today's frontier of foundation model development reflects a computational threshold of 10^26 floating point operations (or "FLOPs"), which in turn "captures only highly resourced developers spending hundreds of millions of dollars to develop foundation models" the bill defines covered models based on the developer's resources and then inappropriately places the burden on developers to justify why models that are above these thresholds should *not* be included in safety protocols.

While the Governor's report discusses foundational models, it is clear that the focus should be on foundation models with such advanced capabilities that they pose novel and potentially severe risks. We urge **SB 53** to be more focused on risk so as to not capture all big foundational models.

Consistent with our position in SB 1047, equating model size to risk manages to make the bill simultaneously overly broad and too narrow as smaller and/or less performant models can present much greater risks than large/higher performant ones. Many 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 a Working Group on Frontier AI, which has since validated such concerns in their Final Reports, finding that small companies may create powerful models that pose safety risks. Yet inexplicably, such models get excluded here. As a result, the bill both fails to adequately address the very real risks posed by small but malicious models and imposes significant costs on innovating performant but responsible ones.

There are two concerns that relying exclusively on FLOPs raises. First, it is a flawed proxy for catastrophic risks that ignores a model's capabilities. Second, it risks becoming obsolete in a short period of time requiring the law to change yet again. The Final Report also noted the variety of reasons why FLOPs are an unreliable proxy for risk and recommended using them just for screening models *if used at all,* saying "we conclude that if training compute thresholds are used at all, they may function best when used as an initial filter to cheaply screen for entities that may warrant greater scrutiny." The report was also critical of other proxies related to size, such as money spent on computing power for model development – a secondary threshold attempted in SB 1047 which we would caution against as a fallback here.

Adaptability is important—as is the Legislature's authority over major policy decisions

The Governor's veto of SB 1047 emphasized the importance of adaptability—a theme echoed in the Governor's Working Group Final Report, which states that "[m]ost fundamentally, the pace of technological change will require adaptive approaches that allow for thresholds to be flexible and readily updated to avoid ossification." We appreciate that **SB 53**, as currently drafted, attempts to incorporate a mechanism for adaptability by allowing the scope of the law to evolve as science progresses. Once again, however, its success is limited by tethering that adaptability to developer size rather than model risk, and by limiting the role of scientific evidence to which the thresholds can adapt.

Specifically, the bill authorizes the AG to annually revise the definition of a "large developer" to ensure it captures "well-resourced large developers at the frontier of artificial intelligence development." (*See* proposed Sec. 22757.15(a).) As noted above, this approach is based on a presumption that size or resources are a sufficient proxy for risk or capability, and in doing so, it disregards the possibility that a risk- or capability-based framework may be more appropriate for determining scope.

Comparatively, proposed Section 22757.15(c) establishes a different mechanism for addressing risk from smaller or less well-resourced developers. If the AG determines that such developers, or those significantly behind the frontier of AI, may nonetheless pose a substantial catastrophic risk, the AG must submit a detailed report to the Legislature along with a proposal for addressing that risk. Yet the AG is expressly prohibited from unilaterally expanding the "large developer" definition to include those entities without legislative approval.

This distinction reflects a tension at the heart of the bill. On one hand, in legislation aimed at promoting safety, it is essential to have mechanisms that can respond quickly to rapid advancements in model capabilities, particularly those nearing thresholds associated with catastrophic risk. On the other hand, it is equally important to avoid overreach – whether by stifling innovation in lower-resourced environments or by granting non-legislative bodies (particularly ones without subject matter expertise) the authority to redefine the law's scope without adequate public input or legislative oversight. By no means should the AG be given unfettered or unchecked authority to make decisions as to who is and is not subject to this law. Regardless, a genuinely adaptive or iterative regulatory framework responds to all relevant scientific evidence, not some of it.

Furthermore, we note that granting the AG *authorization* to *annually* pass regulations does not have a different impact than merely granting the AG authorization to pass regulations. The AG could just as easily decide to examine these issues and pass regulations every five years. There is nothing ***requiring*** the AG to examine the evidence annually. Arguably, an annual review each interim by a legislative oversight committee would provide greater assurance of an adaptive definition than such an authorization, and it would do so without giving the AG such unfettered authority and without giving away legislative authority. That is not to say that the committee oversight process would be the appropriate alternative given all factors to be considered; simply that **SB 53** has not yet found the right approach. Anticipating the development of national or international benchmarks for evaluating model capabilities in the text could be an alternative approach supported by the Working Group's final report.

The definitions of catastrophic risk, critical safety incidents, and dangerous capabilities are vague and contradictory

> **SB 53** relies on three terms that appear to include unique characterizations of risks for which developers must evaluate and mitigate against. Unfortunately, the terms cover similar grounds, making use of these three separate yet similar terms is both confusing and overly complex. For example, *catastrophic risk* is defined via thresholds for injuries or death to persons and dollars of loss resulting from *dangerous capabilities*. This creates concern given the inclusion of widely available capabilities from existing software tools, such as the ability to "assist in a cyberattack" or commit a crime with "limited human intervention" within the definition of "dangerous capabilities". While *dangerous capabilities* are only relevant to the extent they create *catastrophic risk*, as described above, "catastrophic risk" doesn't focus on risk of harm posed by unique risks associated with advanced AI capabilities.

It appears to include currently existing risks that AI models of all sizes may be misused in ways that can cause multiple deaths and high property damage, much like the internet and computers can be used to disrupt critical infrastructures with ransomware attacks that create similar risks of bodily injury, death, and financial losses and damage.

Next, there is also an overbreadth issue, starting with the proposed definition of "catastrophic risk" which requires that the harm will materially contribute in death or serious injury to more than 100 people or more than one billion dollars in damage to rights in money or property arising from a single incident, scheme, or course of conduct involving a dangerous capability. The definition is contradictory, listing both single incident, as well as scheme or course of conduct, vastly expanding the scope of the bill. It is unnecessarily and detrimentally overly broad because a focused SSP is more effective and would ideally be focused on certain categories of risks that exclude things such as blackmail, theft by deception, federation operation, publicly available data, nonmaterial contribution, or cyber threats for example.

Finally, as currently drafted, **SB 53's** overall "critical safety incidents" approach is fundamentally incompatible with the way the AI ecosystem, for both open-source and closed models, as it would impose reporting obligations regarding downstream developer incidents that are infeasible for a developer who does not control the AI system involved in the incident. We note that whereas the Governor's Work Group report recognized the full AI

ecosystem value chain, **SB 53** still needs to more 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 open-source models, eliminate vague triggers, and provide flexibility in reporting timelines to accommodate investigation.

Ultimately, each of the definitions need reworking to focus on advanced capabilities of frontier models and the unique risks they may pose. While doing so, we suggest that an effort be made to unify the terminology and substantive criteria as much as possible to align to the focus recommended in the Final Report. Most important, we feel it critical that **SB 53** better align with the marginal risk standard adopted in the report. As framed by the Governor's Work Group: "that policymakers 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?"

Required disclosures are far too specific, which poses significant security risks and exploitation by bad actors

As currently drafted, proposed Section 22757.12(a) requires that a large developer "write, implement, and clearly and conspicuously publish on its internet website a safety and security protocol that describes in specific detail" specified information detailing the testing procedures that the large developer uses to assess catastrophic risks from its foundation model, among other things.

We have significant concerns about describing many of the required elements in "detail." First, "describing in detail" certain items in a "protocol," such as "testing procedures," "mitigations," and whether evaluations are reproducible are elements of a protocol or development framework that should be adaptable to future scientific developments and flexible based on the capabilities displayed of a given model. A developer should not become beholden to use specific mitigation techniques because they are in their published protocol if a better more effective mitigation is available. Whether capabilities associated with catastrophic risks are reproducible may depend on the model, and the risk was detected.

But also, required disclosures should allow for a level of generality to protect models from exploitation and attack. For example, the requirement in proposed subdivision (a)(6) that a developer describe in detail the "cybersecurity practices and how the large developer secures unreleased model weights from unauthorized modification or transfer by internal or external parties," creates security vulnerabilities for the developer by unnecessarily disclosing detailed information about their cyber protections aiding would be attackers. This is almost akin to a bank putting a blueprint of the location of its security cameras and the protocols of their security guards, and the frequency with which the vault is emptied, online. Once again, as noted at the outset of this letter, the Final Report advised caution for such reasons, stating:

> 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.

Whistleblower provisions

The whistleblower protection chapter needs additional protection against abuse, which could have an impact on a developer's trade secrets, intellectual property, and reputation. Limiting this protection to those involved in critical risk as well as corporate officers helps to mitigate abuse, but would still capture a potentially large group of people given that are "involved with assessing, managing, or addressing critical risk" which could include anyone who works on the trust and safety and testing of an AI model, not just those working at the highest level of assessment of critical risk, and given the size of the companies in scope, there's a potential for abuse. There are different mechanisms to mitigate this risk which we are exploring.

Also, **SB 53** expands "employee" to include contractors and independent workers. The definition should not be changed to include non-employees like contractors. This creates unnecessary risks of misclassification later. The bill could achieve the same stated end to create a definition for the bill of a "covered person" but that includes both employees and contractors as covered individuals.

Additional considerations, including areas of competing viewpoints in industry: something the Legislature must take seriously to avoid granting competitive advantages

There are other considerations as a result of the July 8th amendments that include the following:

- **SSP/protocol requirements lack an intent standard for materially false or misleading statements** (*see* proposed Section 22757.12(e)): the prohibition on false and misleading statements needs an intent standard such as "intentionally", or at least "intentionally or recklessly". Note: whereas *knowingly* infers an awareness of one's action, *intentionally* refers to a conscious objective to cause a result (*i.e.*, something done with reason and purpose).
- **Incident reporting specificity and timeframe:** proposed Section 22757.13(b) requires reporting within 15 days but does not provide flexibility for 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.
- **Upstream versus downstream developers:** we greatly appreciate the effort to focus on transparency and liability in **SB 53,** compared to SB 1047. However, it still needs to better delineate responsibilities between upstream and downstream developers. Developers who pretrain models should not be held liable for fine-tuning or modifications by downstream developers. Clarification is needed to ensure the initial developer is not treated as the developer of a subsequently modified model.
- **Transparency report**: the requirement for a new "transparency report" for every new foundation model should be eliminated. Instead, the bill should allow developers to rely on existing transparency practices such as model cards, requiring only a high-level summary of evaluations and the involvement of third-party assessments, if any.
- **Confidentiality**: we hope to see redactions be broadened beyond trade secrets and cybersecurity information, for example to cover other confidential or proprietary information.

Finally, with respect to enforcement, we ask that the bill 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.

Again, while we understand your focus on the issue of AI models and hope to continue working with you on the issue, the correct approach for California, given the immense promise of this technology, would be to focus on actual risks and harms, and to provide appropriate and secure disclosures. As drafted, however, for all the aforementioned reasons, we must **OPPOSE UNLESS AMENDED SB 53 (Wiener)**.

REGISTERED SUPPORT / OPPOSITION:

Support

Ai for Animals
Ai Futures Project
Ai Lab Watch
Ai Policy Tracker
All Girls Allowed
Apart Research
Association for Long Term Existence and Resilience (ALTER)
Berkeley Existential Risk Initiative (BERI)
Center for Ai and Digital Policy
Center for Ai Policy
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
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
Indivisible California Statestrong
Momentum
Nonlinear
Oakland Privacy
Redwood Research
Secure Ai Future
Secure Ai Project
Seiu Califonia

Tech Oversight California
Techequity Action
The Brandes Lab At Nyu
The Midas Project
The Signals Network
Trevi Digital Assets Fund
University of California

Oppose

Chamber of Progress
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