Date of Hearing: June 18, 2024

ASSEMBLY COMMITTEE ON PRIVACY AND CONSUMER PROTECTION
Rebecca Bauer-Kahan, Chair
SB 1047 (Wiener) – As Amended June 5, 2024

AS PROPOSED TO BE AMENDED

**SENATE VOTE**: 32-1

**SUBJECT**: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act

## SYNOPSIS

> *"We are current and former employees at frontier AI companies, and we believe in the potential of AI technology to deliver unprecedented benefits to humanity.*
>
> *We also understand the serious risks posed by these technologies. These risks range from the further entrenchment of existing inequalities, to manipulation and misinformation, to the loss of control of autonomous AI systems potentially resulting in human extinction. AI companies themselves have acknowledged these risks, as have governments across the world and other AI experts.*
>
> *We are hopeful that these risks can be adequately mitigated with sufficient guidance from the scientific community, policymakers, and the public. However, AI companies have strong financial incentives to avoid effective oversight, and we do not believe bespoke structures of corporate governance are sufficient to change this."*

*The above paragraphs appear at the start of an open letter titled "A Right to Warn about Advanced Artificial Intelligence," released on June 2, 2024 by a group of current and former OpenAI employees. The letter calls on AI companies to commit to various principles of openness and transparency, highlighting these companies' "weak obligations" to share their knowledge of AI's risks with the world.*

*This bill seeks to strengthen those obligations in order to mitigate the risk of catastrophic harms from AI models so advanced that they are not yet known to exist. SB 1047, as proposed to be amended, would require the developers of such models – which cost at least $100 million in computing power to train – to create good governance programs before initiating training. Following training, developers would be required to perform risk assessments on their models, subject to third party auditing, before using or releasing them. The bill creates a Division of Frontier Models in the Government Operations Agency to oversee this process. The bill also adds whistleblower protections; requires operators of computer clusters to implement "know your customer" requirements, including the ability to shut down any resources being used to train an advanced AI model; and creates a public computing cluster known as "CalCompute" in the Department of Technology. The Attorney General is charged with enforcing the bill's requirements.*

*Proposed Committee amendments clarify and strengthen the bill's provisions by, among other things: eliminating the limited duty exemption; adjusting the structure of the Frontier Model*

*Division and placing it under the Board of Frontier Models; enabling the Frontier Model Division to update the definition of "covered model" beginning in 2027; requiring third party audits beginning in 2028; and clarifying the bill's language and scope. The full text of the bill, as proposed to be amended, is included at the end of the analysis.*

*The net result is a strong framework for effective oversight in the face of "the loss of control of autonomous AI systems potentially resulting in human extinction." Moving forward, the devil will be in the details. Certain issues require elaboration and refinement before the bill can be successfully implemented. The analysis details commitments by the author to continue working with the Committee on future amendments relating to enforcement provisions and whistleblower protections, positioning CalCompute within California's university system, fleshing out the structure and responsibilities of the Board of Frontier Models, and further refinements to the bill's terminology.*

*This bill has generated a great deal of commentary, consternation, and misconception. To set the record straight: SB 1047's requirements are not onerous. For one, they only impact models that cost over $100 million in computing power to train – this threshold is baked into the definition of "covered model" provided by the bill and cannot be altered except by future legislative action. Secondly, the risk assessments developers would be required to perform broadly align with national and international guidelines, as well as procedures these developers already claim to implement. Finally, the bill only requires that developers implement shutdown capabilities for rogue models they themselves control. The open source community can rest easy knowing the models they download will not contain an immutable kill switch.*

*This bill does not create a state-sponsored licensing regime for AI; it does not ban the creation and use of AI above a certain compute threshold; and it does not create exorbitant costs for startups seeking to train large models. Performing basic risk assessments before using and releasing powerful, generally-capable models – and prohibiting their use only in extreme cases involving unreasonable risks of mass casualties or massive economic damages – is the bare minimum that Californians should expect of an industry claiming to have their best interests at heart.*

*This bill is sponsored by Center for AI Safety Action Fund, Economic Security California Action, and Encode Justice. It is supported by a variety of advocacy groups including the Future Society. It is opposed by a variety of industry trade associations including the California Chamber of Commerce, Technet, and the Chamber of Progress. If the bill passes this Committee, it will next be heard by the Assembly Judiciary Committee.*

**SUMMARY:** Requires developers of especially advanced AI systems conduct risk assessments of those systems, to be overseen by the Frontier Model Division in the Government Operations Agency. Specifically, **this bill**:

1) Defines "artificial intelligence" 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.

2) Defines "fine-tuning" to mean adjusting the model weights of a trained model by exposing it to additional data.

3) Before Jan. 1, 2027, defines "covered model" to mean either of the following:

a) An artificial intelligence model trained using a quantity of computing power greater than 10^26 integer or floating-point operations, the cost of which exceeds $100 million dollars when calculated using the average market price of cloud compute at the start of training as reasonably assessed by the developer.

b) An artificial intelligence model created by fine-tuning a covered model using a quantity of computing power equal to or greater than 3 times 10^25 integer or floating-point operations.

4) On and after Jan. 1, 2027, defines "covered model" to mean either of the following:

a) An artificial intelligence model trained using a quantity of computing determined by the Frontier Model Division, the cost of which exceeds $100 million dollars when calculated using the average market price of cloud compute at the start of training as reasonably assessed by the developer.

b) An artificial intelligence model created by fine-tuning a covered model using a quantity of computing power that exceeds a threshold determined by the Frontier Model Division.

5) Defines "post-training modification" to mean modifying the capabilities of a covered model by any means, including, but not limited to, fine-tuning, providing the model with access to tools or data, removing safeguards against hazardous misuse or misbehavior of the model, or combining the model with, or integrating it into, other software.

6) Defines "covered model derivative" to mean any of the following:

a) An unmodified copy of a covered model.

b) A copy of a covered model that has been subjected to post-training modifications unrelated to fine-tuning.

c) Before January 1, 2027, a copy of a covered model that has been fine-tuned using a quantity of computing power not exceeding 3 times 10^25 integer or floating-point operations.

d) On and after January 1, 2027, a copy of a covered model that has been fine-tuned using a quantity of computing power not exceeding a threshold determined by the Frontier Model Division.

7) Defines "developer" to mean a person that performs the initial training of a covered model either by training a model using a sufficient quantity of computing power, or by fine-tuning an existing model using a sufficient quantity of computing power.

8) Defines "critical harm" to mean any of the following:

a) The creation or use of a chemical, biological, radiological, or nuclear weapon in a manner that results in mass casualties.

b) Mass casualties or at least $500 million of damage resulting from cyberattacks on critical infrastructure, occurring either in a single incident or over multiple related incidents.

c) Mass casualties, or at least $500 million of damage resulting from an artificial intelligence model autonomously engaging in conduct that would constitute a serious or violent felony under the Penal Code if were undertaken by a human with the requisite mental state.

d) Other grave harms to public safety and security that are of comparable severity to the listed harms.

9) Defines "full shutdown" to mean the cessation of operation of any of the following:

a) The training of a covered model.

b) A covered model.

c) All covered model derivatives controlled by a developer.

10) Requires that before a developer initially trains a covered model, they do all of the following:

a) Implement protections to prevent unauthorized access, misuse, or unsafe post-training modification of the covered model and all associated covered model derivatives controlled by the developer.

b) Implement the capability to enact a full shutdown.

c) Implement a written safety and security protocol that does all of the following:

1. Provides reasonable assurance that the developer will not produce a covered model or covered model derivative that poses an unreasonable risk of causing or enabling a critical harm.

2. Identifies specific tests that would be sufficient to provide reasonable assurance that covered models and covered model derivatives do not pose an unreasonable risk of causing or enabling a critical harm.

3. Describes the conditions under which a developer would enact a full shutdown.

4. Describes the procedure by which the safety and security protocol may be modified.

d) Implement the safety and security protocol and designate senior personnel to be responsible for ensuring compliance, including by conducting third party audits.

e) Conduct an annual review of the safety and security protocol and make modifications as necessary.

f) Provide a copy of the safety and security protocol to the Frontier Model Division when it is initially created, and within 10 business days of any modifications.

11) Requires that before using a covered model or covered model derivative, or making a covered model or covered model derivative available for commercial or public use, the developer do all of the following:

a) Assess whether the covered model is reasonably capable of causing a critical harm.

b) Implement reasonable safeguards to prevent a covered model or associated covered model derivatives from causing a critical harm.

c) Ensure, to the extent reasonably possible, that the actions of covered models and covered model derivatives can be attributed to them.

d) Beginning Jan. 1, 2028, obtain a certificate of compliance from a third party auditor who has been accredited by the Frontier Model Division.

12) Prohibits a developer from using a covered model or covered model derivative commercially or publicly, or making a covered model or covered model derivative available for commercial or public use, if there is an unreasonable risk that the covered model or covered model derivative can cause or enable a critical harm.

13) Requires a developer of covered model to annually submit a certification of compliance to the Frontier Model Division for as long as the covered model or any covered model derivatives controlled by the developer are used commercially or publicly, or remain available for commercial or public use.

14) Defines "artificial intelligence safety incident" to mean an incident that demonstrably increases the risk of a critical harm occurring by means of any of the following:

a) A covered model autonomously engaging in behavior other than at the request of a user.

b) Theft, misappropriation, malicious use, inadvertent release, unauthorized access, or escape of the model weights of a covered model.

c) The critical failure of technical or administrative controls, including controls limiting the ability to modify a covered model.

d) Unauthorized use of a covered model to cause or enable a critical harm.

15) Requires a developer of a covered model to report each artificial intelligence safety incident to the Frontier Model Division within 72 hours of the developer learning of the safety incident.

16) Defines "computing cluster" to mean a set of machines transitively connected by data center networking of over 100 gigabits per second that has a theoretical maximum computing capacity of at least 10^20 integer or floating-point operations per second and can be used for training artificial intelligence.

17) Requires a person that operates a computing cluster to implement written policies and procedures to do all of the following when a customer utilizes compute resources sufficient to train a covered model:

a) Obtain a prospective customer's basic identifying information and business purpose for utilizing the computing cluster.

b) Assess whether the customer intends to train a covered model.

c) Retain the customer's internet protocol addresses used to access the cluster, along with the time and date of access or administrative action.

d) Maintain the above records for 7 years and provide them to the Frontier Model Division or the Attorney General upon request.

e) Implement the capability to enact a full shutdown of any resources being used to train a covered model.

18) Requires a developer of a covered model that provides commercial access to it to provide a transparent, uniform, publicly available price schedule for the purchase of access to that model at a given level of quality and quantity subject to the developer's terms of service and prohibits developers from engaging in unlawful discrimination or noncompetitive activity in determining price or access. Operators of computing clusters are required to do the same with respect to computing clusters. However, a person that operates a computing cluster may provide free, discounted, or preferential access to public entities, academic institutions, or for noncommercial research purposes.

19) Authorizes the Attorney General, if they have reasonable cause to believe that a person is in violation of these provisions, to bring an action seeking recovery of preventive relief, including a permanent or temporary injunction, restraining order, or other order against the person responsible for the violation, including deletion of the covered model and the weights utilized in that model. Monetary damages to persons aggrieved and a court order for a full shutdown are also available. However, these remedies are only available in response to harm or an imminent risk or threat to public safety. The Attorney General may also recover a civil penalty in an amount not exceeding 10 percent of the cost, excluding labor, to develop the covered model for a first violation and in an amount not exceeding 30 percent of the cost, excluding labor, to develop the covered model for any subsequent violation.

20) Subjects liable defendants to joint and several liability and instructs the court to disregard corporate formalities under specific conditions:

a) Where steps were taken in the development of the corporate structure among affiliated entities to purposely and unreasonably limit or avoid liability.

b) Where the corporate structure of the developer or affiliated entities would frustrate recovery of penalties or injunctive relief.

21) Prohibits a developer from preventing an employee from disclosing information to the Attorney General if the employee has reasonable cause to believe that the information indicates that the developer is out of compliance. A developer shall not retaliate against an employee for disclosing such information. Developers must provide clear notice to all employees working on covered models of their rights and responsibilities under this section. The Attorney General may publicly release any complaint, or a summary of that complaint, if disclosure will serve the public interest.

22) Clarifies that the duties and obligations imposed 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.

23) Establishes the Board of Frontier Models in the Government Operations Agency. Allows the Governor to appoint an executive director of the Board, subject to Senate confirmation, to exercise all duties and functions necessary to ensure that the responsibilities of the board are successfully discharged. Specifies that the board shall be composed of 5 members, as follows:

    a) A member of the open-source community, appointed by the Governor, subject to Senate confirmation.

    b) A member of the artificial intelligence industry, appointed by the Governor, subject to Senate confirmation.

    c) A member of academia, appointed by the Governor, subject to Senate confirmation.

    d) A member appointed by the Speaker of the Assembly.

    e) A member appointed by the Senate Rules Committee.

24) Establishes the Frontier Model Division in the Government Operations Agency, under the direct supervision of the Board. Requires the Division to do the following:

    a) Annually review certifications received from developers.

    b) Advise the Attorney General on potential violations of the bill's provisions.

    c) Establish an accreditation process for third party auditors.

    d) Publish anonymized safety reports.

    e) Issue guidance describing categories of AI safety events likely to constitute a state of emergency.

    f) Appoint and consult with an advisory committee for open-source AI, which shall:

        1. Levy fees, including an assessed fee for the submission of a certification, in an amount sufficient to cover the reasonable costs of administering the Frontier Model Division's responsibilities.

        2. Develop and submit to the Judicial Council proposed model jury instructions for actions involving violations related to developers of covered models.

        3. On or before Jan 1, 2027, and annually thereafter, issue regulations to update the definition of a "covered model" to ensure that it accurately reflects technological developments, scientific literature, and widely-accepted national and international standards and applies to artificial intelligence models that pose the greatest risk of enabling critical harms. The updated definition shall contain the following:

            i. The initial compute threshold that an artificial intelligence model must exceed to be considered a covered model.

ii. The fine-tuning compute threshold that an artificial intelligence model must meet to be considered a covered model.

25) Tasks the Department of Technology with creating a public cloud computing cluster known as CalCompute through the commissioning of consultants with specified objectives, first of which is to study the safe and secure deployment of large-scale AI models. The consultants shall include representatives of national laboratories, public universities, and any relevant professional associations or private sector stakeholders. They shall evaluate and incorporate the following considerations into their plan:

a) An analysis of the public, private, and nonprofit cloud platform infrastructure ecosystem, including, but not limited to, dominant cloud providers, the relative compute power of each provider, the estimated cost of supporting platforms as well as pricing models, and recommendations on the scope of CalCompute.

b) The process to establish affiliate and other partnership relationships to establish and maintain an advanced computing infrastructure.

c) A framework to determine the parameters for use of CalCompute, including, but not limited to, a process for deciding which projects will be supported by CalCompute and what resources and services will be provided to projects.

d) A process for evaluating appropriate uses of the public cloud resources and their potential downstream impact, including mitigating downstream harms in deployment.

e) An evaluation of the landscape of existing computing capability, resources, data, and human expertise in California for the purposes of responding quickly to a security, health, or natural disaster emergency.

f) An analysis of the state's investment in the training and development of the technology workforce, including through degree programs at the University of California, the California State University, and the California Community Colleges.

g) A process for evaluating the potential impact of CalCompute on retaining technology professionals in the public workforce.

26) Authorizes the Department of Technology to receive private donations, grants, and local funds, in addition to allocated funding in the annual budget, to effectuate the establishment of CalCompute.

**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.)

**FISCAL EFFECT**: As currently in print this bill is keyed fiscal.

**COMMENTS**:

**1) 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.

**2) Risk management.** According to the Senate Judiciary Committee's comprehensive analysis of the bill:

> With recent dramatic advances in the capabilities of AI systems, the need for frameworks for accountability and responsible development have become ever more urgent.
>
> In January of 2017, AI researchers, economists, legal scholars, ethicists, and philosophers met in Asilomar, California to discuss principles for managing the responsible development of AI. The collaboration resulted in the Asilomar Principles. Aspirational rather than prescriptive, these 23 principles were intended to initiate and frame a dialogue by providing direction and guidance for policymakers, researchers, and developers. Its endorsers include 1,200 leading experts in the field of AI, including DeepMind founder Demis Hassabis and the late Stephen Hawking.[1]
>
> The Legislature subsequently adopted ACR 215 (Kiley, Ch. 206, Stats. 2018), which added the State of California to that list by endorsing the Asilomar Principles as guiding values for

[1] Future of Life Institute, "Asilomar AI Principles," Aug. 11, 2017, https://futureoflife.org/open-letter/ai-principles/.

the development of artificial intelligence and related public policy. In broad strokes, those principles aim to do the following:

- Research issues: create beneficial AI; direct funding toward beneficial innovation; maintain constructive and healthy exchanges between AI researchers and policymakers; promote a culture of trust, cooperation, and transparency among researchers and developers of AI; and avoid corner-cutting on safety standards.

- Ethics and values: promote safety, failure transparency, judicial transparency, and responsible innovation; align human values with innovation; protect privacy and liberty; ensure that the benefits and prosperity created by AI are broadly shared; maintain human control over AI; develop AI that supports rather than subverts social and civil processes; and avoid an AI arms race.

- Longer-term issues: avoid assumptions regarding the capabilities of AI; give AI its due attention; and recognize that its risks are potentially catastrophic or existential.

As directed by the National AI Initiative Act of 2020, NIST developed the AI Risk Management Framework to assist entities designing, developing, deploying, and using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems. That framework highlights the serious risks at play and the uniquely challenging nature of addressing them in this context:

> Artificial intelligence (AI) technologies have significant potential to transform society and people's lives – from commerce and health to transportation and cybersecurity to the environment and our planet. AI technologies can drive inclusive economic growth and support scientific advancements that improve the conditions of our world. AI technologies, however, also pose risks that can negatively impact individuals, groups, organizations, communities, society, the environment, and the planet. Like risks for other types of technology, AI risks can emerge in a variety of ways and can be characterized as long- or short-term, high or low-probability, systemic or localized, and high- or low-impact.
>
> While there are myriad standards and best practices to help organizations mitigate the risks of traditional software or information-based systems, the risks posed by AI systems are in many ways unique. AI systems, for example, may be trained on data that can change over time, sometimes significantly and unexpectedly, affecting system functionality and trustworthiness in ways that are hard to understand. AI systems and the contexts in which they are deployed are frequently complex, making it difficult to detect and respond to failures when they occur. AI systems are inherently socio-technical in nature, meaning they are influenced by societal dynamics and human behavior. AI risks – and benefits – can emerge from the interplay of technical aspects combined with societal factors related to how a system is used, its interactions with other AI systems, who operates it, and the social context in which it is deployed.
>
> These risks make AI a uniquely challenging technology to deploy and utilize both for organizations and within society. [. . .]
>
> AI risk management is a key component of responsible development and use of AI systems. Responsible AI practices can help align the decisions about AI system design,

development, and uses with intended aim and values. Core concepts in responsible AI emphasize human centricity, social responsibility, and sustainability. AI risk management can drive responsible uses and practices by prompting organizations and their internal teams who design, develop, and deploy AI to think more critically about context and potential or unexpected negative and positive impacts. Understanding and managing the risks of AI systems will help to enhance trustworthiness, and in turn, cultivate public trust.[2]

More recently the Biden Administration has published its Blueprint for an AI Bill of Rights, which is a set of five principles and associated practices to help guide the design, use, and deployment of AI to protect the rights of the American public:

- Safe and Effective Systems: You should be protected from unsafe or ineffective systems. Automated systems should be developed with consultation from diverse communities, stakeholders, and domain experts to identify concerns, risks, and potential impacts of the system.

- Algorithmic Discrimination Protections: Designers, developers, and deployers of automated systems should take proactive and continuous measures to protect individuals and communities from algorithmic discrimination and to use and design systems in an equitable way. This protection should include proactive equity assessments as part of the system design, use of representative data and protection against proxies for demographic features, ensuring accessibility for people with disabilities in design and development, pre-deployment and ongoing disparity testing and mitigation, and clear organizational oversight.

- Data Privacy: You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used. You should be protected from violations of privacy through design choices that ensure such protections are included by default, including ensuring that data collection conforms to reasonable expectations and that only data strictly necessary for the specific context is collected. Designers, developers, and deployers of automated systems should seek your permission and respect your decisions regarding collection, use, access, transfer, and deletion of your data in appropriate ways and to the greatest extent possible; where not possible, alternative privacy by design safeguards should be used. Systems should not employ user experience and design decisions that obfuscate user choice or burden users with defaults that are privacy invasive. Consent should only be used to justify collection of data in cases where it can be appropriately and meaningfully given. Any consent requests should be brief, be understandable in plain language, and give you agency over data collection and the specific context of use; current hard-to-understand notice-and-choice practices for broad uses of data should be changed. Enhanced protections and restrictions for data and inferences related to sensitive domains, including health, work, education, criminal justice, and finance, and for data pertaining to youth should put you first. In sensitive domains, your data and related inferences should only be used for necessary functions, and you should be protected by ethical review and use prohibitions. You and your

[2] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework," Jan. 2023, https://doi.org/10.6028/NIST.AI.100-1.

communities should be free from unchecked surveillance; surveillance technologies should be subject to heightened oversight that includes at least pre-deployment assessment of their potential harms and scope limits to protect privacy and civil liberties. Continuous surveillance and monitoring should not be used in education, work, housing, or in other contexts where the use of such surveillance technologies is likely to limit rights, opportunities, or access. Whenever possible, you should have access to reporting that confirms your data decisions have been respected and provides an assessment of the potential impact of surveillance technologies on your rights, opportunities, or access.

- Notice and Explanation: You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you. Designers, developers, and deployers of automated systems should provide generally accessible plain language documentation including clear descriptions of the overall system functioning and the role automation plays, notice that such systems are in use, the individual or organization responsible for the system, and explanations of outcomes that are clear, timely, and accessible. Such notice should be kept up-to-date and people impacted by the system should be notified of significant use case or key functionality changes. You should know how and why an outcome impacting you was determined by an automated system, including when the automated system is not the sole input determining the outcome.

- Human Alternatives, Consideration, and Fallback: You should be able to opt out from automated systems in favor of a human alternative, where appropriate. Appropriateness should be determined based on reasonable expectations in a given context and with a focus on ensuring broad accessibility and protecting the public from especially harmful impacts.[3]

TechEquity, an organization committed to ensuring technology's evolution benefits everyone equitably, has also laid out their straightforward AI Policy Principles:

- People who are impacted by AI must have agency to shape the technology that dictates their access to critical needs like employment, housing, and healthcare.

- The burden of proof must lie with developers, vendors, and deployers to demonstrate that their tools do not create harm—and regulators, as well as private [individuals], should be empowered to hold them accountable.

- Concentrated power and information asymmetries must be addressed in order to effectively regulate the technology.[4]

The need for thoughtful regulation and accountability is especially urgent with regard to the existential risks that many believe unfettered AI advancement poses. In response to these

---

[3] The White House, "Blueprint for an AI Bill of Rights," Oct. 2023, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.
[4] TechEquity, "AI Policy Principles," Mar. 2024, https://techequitycollaborative.org/wp-content/uploads/2024/03/AI_Policy_Principles.pdf.

risks, the Future of Life Institute published an open letter early last year, calling for a pause on giant AI experiments:

> Contemporary AI systems are now becoming human-competitive at general tasks, and we must ask ourselves: Should we let machines flood our information channels with propaganda and untruth? Should we automate away all the jobs, including the fulfilling ones? Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? Should we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders. Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable. This confidence must be well justified and increase with the magnitude of a system's potential effects. OpenAI's recent statement regarding artificial general intelligence, states that "At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models." We agree. That point is now.
>
> Therefore, we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4. This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium.[5]

Signatories to the letter include Stuart Russell, Berkeley, Professor of Computer Science, director of the Center for Intelligent Systems, and co-author of the standard textbook "Artificial Intelligence: a Modern Approach"; Elon Musk, CEO of SpaceX, Tesla & X; and Steve Wozniak, Co-founder, Apple.

Subsequent to that letter, the Center for AI Safety released another open letter signed by a wide-ranging group of industry leaders, researchers, and engineers working in AI that highlighted the existential risk posed by unethical AI development and the urgency of the issue. The statement simply read: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[6]

This was signed by the most cited researchers of AI, including Dr. Yoshua Bengio and Dr. Geoffrey Hinton; both Turing Award winners and considered the "godfathers" of modern AI. In addition, prominent executives at the leading AI development companies also signed on, including Ilya Sutskever, co-founder and chief scientist, OpenAI; Sam Altman, chief executive of OpenAI; Demis Hassabis, chief executive of Google DeepMind; and Dario Amodei, chief executive of Anthropic.

While the future is unclear, the need to respond to these potential harms now is evident. The Center for New American Security puts a fine point on it:

> While there is significant uncertainty in how the future of AI develops, current trends point to a future of vastly more powerful AI systems than today's state of the art. The

---

[5] Future of Life Institute, "Pause Giant AI Experiments: An Open Letter," Mar. 22, 2023, https://futureoflife.org/wp-content/uploads/2023/05/FLI_Pause-Giant-AI-Experiments_An-Open-Letter.pdf.
[6] Center for AI Safety, "Statement on AI Risk," May 30, 2023, https://www.safe.ai/work/statement-on-ai-risk#open-letter.

> most advanced systems at AI's frontier will be limited initially to a small number of actors but may rapidly proliferate. Policymakers should begin to put in place today a regulatory framework to prepare for this future. Building an anticipatory regulatory framework is essential because of the disconnect in speeds between AI progress and the policymaking process, the difficulty in predicting the capabilities of new AI systems for specific tasks, and the speed with which AI models proliferate today, absent regulation. Waiting to regulate frontier AI systems until concrete harms materialize will almost certainly result in regulation being too late.

**3) What this bill, as proposed to be amended, would do.** SB 1047 would require developers of "covered models" to conduct risk assessments on those models before using them or making them available for public use. Until 2027, models are considered "covered models" if they are trained using roughly ten times as much computing power as any model that exists today and the costs of that quantity of compute power exceeds $100 million.

Beginning in 2027, the Frontier Model Division may adjust the compute threshold determining which models are considered "covered models." Developers are prohibited from using or releasing covered models that pose an unreasonable risk of causing or enabling "critical harms," resulting in mass casualties or $500 million in damages, as specified. Beginning January 1, 2028, developers of covered models will be subject to third party audits.

The bill requires the operators of computing clusters to know their customers and implement the ability to shut down any hardware used to train a covered model.

The bill creates the Frontier Model Division in the Government Operations Agency, under control of the Frontier Model Board. The Division is charged with various duties related to overseeing the development of advanced AI in the state, including setting thresholds for the definition of "covered model" and creating a third party auditor accreditation program.

The bill creates CalCompute, a public computing cluster located in the Department of Technology.

Finally, the bill also adds whistleblower protections related to covered models and gives the Attorney General civil enforcement authority.

**4) Author's statement.**

> Large-scale artificial intelligence has the potential to produce an incredible range of benefits for Californians and our economy—from advances in medicine and climate science to improved wildfire forecasting and clean power development. It also gives us an opportunity to apply hard lessons learned over the last decade, as we've seen the consequences of allowing the unchecked growth of new technology without evaluating, understanding, or mitigating the risks. SB 1047 does just that, by developing responsible, appropriate guardrails around development of the largest, most powerful AI systems, to ensure they are used to improve Californians' lives, without compromising safety or security.
>
> SB 1047 will also promote the growth of the AI industry by establishing CalCompute, a public AI research cluster that will allow startups, researchers, and community groups to participate in the development of large-scale AI systems. By providing a broad range of

stakeholders with access to the AI development process, CalCompute will help align large-scale AI systems with the values and needs of California communities.

**5) Analysis.** Rather than retreading ground previously covered by Senate Judiciary Committee, this analysis focuses on outlining and explaining the Committee's proposed amendments.

*Updating the definition of "covered model."* In order to be considered a "covered model" under SB 1047, a model must meet one of two conditions:

1. The model is a new model that is initially trained using a quantity of computing power exceeding 10^26 integer or floating-point operations (and costing at least $100 million).

2. The model is an existing model that is fine-tuned (it receives additional training) using a quantity of computing power exceeding 3 times 10^25 integer or floating-point operations.

The most advanced contemporary models are trained using roughly a tenth of the compute required to be considered a "covered model" under this bill.[7] Could ChatGPT4 be capable of causing a "critical harm"? Possible, but unlikely. However, as the training data that inform AI become more useful, and the architectures that underpin these systems become more carefully constructed, it is not unthinkable that a 10^25 model will one day be capable of causing such a harm. If this comes to pass, the definition of "covered model" will need to be updated. The European Union, importantly, sets a comparable threshold at 10^25 integer or floating-point operations.[8]

Alternatively, 10^26 models may end up being completely harmless. Five years from now it may make no sense for models of this size to be rigorously tested while 10^27, 10^28, and 10^29 models exist in the ecosystem. The definition of "covered model" would benefit from receiving an update in this case as well, if for no other reason than to avoid overwhelming the Frontier Model Division.

The fine-tuning threshold outlined in condition 2 has a more serious issue: fine-tuning takes far less compute than initially training a model. Condition 2 requires that a model receive 3 times 10^25 integer or floating-point operations worth of training before becoming a "new" covered model. Upon meeting this threshold, responsibility for the model switches over to the party that fine-tuned the model. This means the original developer is responsible for guaranteeing the safety of a covered model within a fine-tuning threshold of nearly 1/3 the original compute that went into its training.

It is not clear that this is possible. A set of leaked specifications for ChatGPT4 (which may not be entirely reliable, but are likely in the right ballpark) suggest that ChatGPT4 was initially trained using 13 trillion tokens.[9] OpenAI claims that their models' behavior begins to change when fine-tuned using 50-100 training samples, or roughly 200,000 tokens (if each sample is

---

[7] Matthias Plappert, Durk Kingma, Max Chen, Cage Zhong, and Penny Deng, "Thoughts on Llama 3," *Factorial Funds*, Apr. 24, 2024, https://www.factorialfunds.com/blog/thoughts-on-llama-3.
[8] European Union, "EU Artificial Intelligence Act," https://artificialintelligenceact.eu/article/51/.
[9] Yam Peleg, "GPT-4's details are leaked," *Twitter*, Jul, 11, 2023, https://archive.is/2RQ8X#selection-463.1-463.28.

~2048 tokens).[10] There is a clear disconnect here. As it becomes more apparent how much initial compute is required for a model to acquire harmful capabilities, it may also become apparent how much fine-tuning compute allows a safe model to become unsafe.

Committee amendments would create a mechanism to update these thresholds in light of new evidence, through regulations passed by the Frontier Model Division beginning in 2027. Importantly, the Division's discretion in adjusting the thresholds is considerably constrained. In updating the threshold, through regulation, the Division must ensure the update accurately reflects technological developments, scientific literature, and widely-accepted national and international standards, and that it applies to AI that pose the greatest risk of enabling critical harms. Additionally, the Division would not have the ability to alter the "cost" threshold of training a covered model. If the compute cost of training a 10^26 model drops below $100 million in future, and 10^26 models are demonstrated to be broadly capable of causing critical harms, this cost threshold may need to be updated through legislative action.

*Eliminating the "limited duty exemption."* The bill in print contains a mechanism for developers to self-certify that their models possess no harmful capabilities, called the "limited duty exemption." If a model qualifies for one of these "exemptions," it is not subject to any of downstream requirements of the bill. Confusingly, developers are asked to make this assessment before a model has been trained—that is, before it exists. Writing in opposition, the California Chamber of Commerce explains why this puts developers in an impossible position:

> SB 1047 still makes it impossible for developers to actually determine if they can provide reasonable assurance that a covered model does not have hazardous capabilities and therefore qualifies for limited duty exemption because it requires developers to make the determination before they initiate training of the covered model . . . Because a developer needs to test the model by training it in a controlled environment to make determination that a model qualifies for the exemption, and yet cannot train a model until such a determination is made, SB 1047 effectively places developers in a perpetual catch-22 and illogically prevents them from training frontier models altogether.

Furthermore, SB 1047 is predicated on the notion that advanced AI models are inherently risky above a certain compute threshold. It makes little sense to fully exempt a subset of these models from the provisions of this bill. Committee amendments abolish the limited duty exemption. Under this new framework, all developers that train covered models must create governance programs and perform basic risk assessments.

*Clarifying the definition of "developer."* The version of the bill in print defines "developer" to mean a person that creates, owns, or otherwise has responsibility for an artificial intelligence model. It then excludes a number of individuals from this definition: "a third-party machine-learning operations platform, an artificial intelligence infrastructure platform, a computing cluster, an application developer using sourced models, or an end-user of an artificial intelligence model." This creates various loopholes: if OpenAI purchases a computing cluster and exclusively uses that cluster to train ChatGPT5, are they then not considered a developer under this bill? Committee amendments simplify the definition of developer, focusing on persons who are in the position to comply with the bill's requirements.

---

[10] OpenAI, "Introducing GPT-4o: our fastest and most affordable flagship model," https://platform.openai.com/docs/guides/fine-tuning.

*Clarifying the scope of a "full shutdown."* SB 1047's "full shutdown" requirement has been a source of constant consternation for the open-source community. CalChamber explains:

> Under SB 1047, developers must build "full shutdown" capabilities into their models and may be held liable for downstream uses over which they have no control, impeding their ability to open-source their models. Ultimately, liability should rest with the user who intended to do harm, as opposed to automatically defaulting to the developer who could not foresee, let alone block, any and all conceivable uses of a model that might do harm. While recent amendments seemingly seek to narrow what is meant by "full shutdown" capabilities, the exclusions are unnecessarily difficult to interpret as drafted (full shutdown "does not mean the cessation of operation of a covered model to which access was granted pursuant to a license that was not created by the licensor…") and altogether insufficient.

Committee amendments simplify and clarify the definition of "full shutdown" such that the shutdown capability can be implemented into hardware used to train or run a model, rather than the model itself. The amendments also serve to exclude covered model derivatives that are outside of the developer's control.

*Reframing the bill in terms of "risk of causing or enabling a critical harm."* SB 1047 previously alternated between the terms "hazardous capabilities" and "critical harms" when describing the dangers associated with advanced AI. The definition of "critical harm" simply pointed to the definition of "hazardous capability," and the definition of hazardous capability was effectively "the capability of a covered model to be used to enable one of [a list of harms]." Committee amendments simplify the bill's language by couching everything in terms of a covered model's capability to cause or enable various critical harms.

*Adjusting a "critical harm" exemption.* The version of the bill in print bakes the following exemption into the definition of "critical harm/hazardous capability": hazardous capability means "the capability of a covered model to be used to enable any of the following harms **in a way that would be significantly more difficult to cause without access to a covered model**." This exemption is crafted broadly and leads to unintended consequences. For example: an autonomous robot that goes on a stabbing spree might not be covered by this bill, as the creator of the model could just as easily have picked up a knife. According to the author and sponsors, the intent of this exemption is to cover information that a model could provide, but that could easily be found elsewhere (such as online.) Committee amendments narrow the scope of this exemption by focusing on harms caused or enabled by information that a covered model outputs if the information is otherwise publicly accessible. This solution is imperfect: for example, it results in advanced AI effectively inheriting various protections afforded to the internet under Section 230. It also fails to account for situations where information found online originated from another advanced AI. The author has committed to continue to work with the Committee on refining this language.

*Closing a "derivative model" loophole.* The bill in print contains a loophole related to the definition of "derivative model." A derivative model includes "a modified or unmodified copy of an artificial intelligence model," and derivative models are exempted from most of the bill's requirements. However, the definition of derivative model does not specify that this model must be owned or operated by an entity other than the original developer. This means that a developer could train a model, copy/paste it to a different set of hardware (but otherwise leave it unaltered), and be considered exempt from the provisions of the bill due to the copy being a "derivative."

Committee amendments close this loophole by recasting the bill in terms of "covered models and covered model derivatives" and requiring developers to be responsible for both.

*Clarifying the bill's prohibition on the use of models with unreasonable risks of causing critical harms.* The bill in print requires the following:

> Before initiating the commercial, public, or widespread use of a covered model that is not subject to a limited duty exemption, a developer of the nonderivative version of the covered model shall do all of the following . . . *Refrain* from initiating the commercial, public, or widespread use of a covered model if there remains an unreasonable risk that an individual may be able to use the hazardous capabilities of the model, or a derivative model based on it, to cause a critical harm. (Emphasis added.)

This provision would appear to ban the deployment of dangerous covered models. However, it is not entirely clear due to the bill's use of the term "refrain," which the Merriam-Webster dictionary defines as "keep[ing] oneself from doing, feeling, or indulging in something and especially from following a passing impulse." In other words, the bill in print arguably does not ban the deployment of a model – it only requires that a developer try their best not to release it. Committee amendments clarify that a developer is prohibited from using or releasing covered models "if there is an unreasonable risk that the covered model or covered model derivative can cause or enable a critical harm."

*Requiring third-party auditing.* The bill in print directs the Frontier Model Division to create an "optional" accreditation process for third party auditing. Committee amends retain this function of the Frontier Model Division, but make third party auditing mandatory. Beginning January 1, 2028, developers must obtain a certificate of compliance from an accredited third party auditor before they can use or make available covered models or covered model derivatives they have trained.

*Adjusting the structure and placement of the Frontier Model Division.* Committee amendments give the Frontier Model Division two new responsibilities under this bill: accrediting third party auditors, and adjusting the compute thresholds in the definition of "covered model." These changes empower the Division to both determine which entities are subject to the provisions of this bill, and determine how strictly to enforce compliance. To ensure that this process is done with proper transparency, accountability, and public participation, Committee amendments create a "Board of Frontier Models" in the Government Operations Agency to supervise and direct the Division. Five members sit on this board: three Governor appointees, one Senate appointee, and one Assembly appointee. Of the Governor's appointees, all three must receive Senate confirmation. One must originate from academia, one from industry, and one from the open source community. By building various perspectives into the leadership of the Board, this Committee hopes to ensure the Division's work remains fair and free of bias. Committee amendments also move the Division itself from the Department of Technology to the Government Operations Agency.

*Various language changes.* Committee amendments update language throughout SB 1047 to clarify the bill's scope and requirements.

*Author commitments.* In addition to the changes outlined above, the author has agreed to continue working with the Committee on several pieces of SB 1047. These are briefly outlined below:

- Restructuring the enforcement scheme in collaboration with the Assembly Judiciary Committee.

- Expanding whistleblower protections in collaboration with the Assembly Judiciary Committee.

- Positioning CalCompute in California's university system, rather than in the Department of Technology.

- Fleshing out the structure and responsibilities of the Board of Frontier Models.

- Workshopping a narrow exemption for "critical harms" that relates to the ability of these models to produce information.

- Adjusting language to clarify the bill's provisions and bring various terminology in line with industry standards, without affecting the overall impact of the bill.

**6) Full text as proposed to be amended.**

CHAPTER 22.6. Safe and Secure Innovation for Frontier Artificial Intelligence Models

22602. As used in this chapter:

(a) "Advanced persistent threat" means an adversary with sophisticated levels of expertise and significant resources that allow it, through the use of multiple different attack vectors, including, but not limited to, cyber, physical, and deception, to generate opportunities to achieve its objectives that are typically to establish and extend its presence within the information technology infrastructure of organizations for purposes of exfiltrating information or to undermine or impede critical aspects of a mission, program, or organization or place itself in a position to do so in the future.

(b) "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, ~~infers,~~***infer*** from the input it receives~~,~~ how to generate outputs that can influence physical or virtual environments ~~and that may operate with varying levels of autonomy~~.

(c) "Artificial intelligence safety incident" means ***an incident that demonstrably increases the risk of a critical harm occurring by means of*** any of the following:

(1) A covered model autonomously engaging in behavior other than at the request of a user ~~that materially increases the risk of a hazardous capability being used~~.

(2) Theft, misappropriation, malicious use, inadvertent release, unauthorized access, or escape of the model weights of a covered model ~~that is not the subject of a limited duty exemption~~.

(3) The critical failure of technical or administrative controls, including controls limiting the ability to modify a covered model ~~that is not the subject of a limited duty exemption~~.

(4) Unauthorized use of ~~the hazardous capability of~~ a covered model ***to cause or enable a critical harm***.

(d) "Computing cluster" means a set of machines transitively connected by data center networking of over 100 gigabits per second that has a theoretical maximum computing capacity of at least 10^20 integer or floating-point operations per second and can be used for training artificial intelligence.

(e) "Covered ~~guidance~~***model***" means either of the following:

~~(1) Guidance issued by the National Institute of Standards and Technology and by the Frontier Model Division that is relevant to the management of safety risks associated with artificial intelligence models that may possess hazardous capabilities.~~

~~(2) Industry best practices, including safety practices, precautions, or testing procedures undertaken by developers of comparable models that are relevant to the management of safety risks associated with artificial intelligence models that may possess hazardous capabilities.~~

~~(f) (1) "Covered model" means an~~***(1) Before January 1, 2027, "covered model" means either of the following:***

***(A) An*** artificial intelligence model ~~that was~~ trained using a quantity of computing power greater than 10^26 integer or floating-point operations, ~~and~~ the cost of ~~that quantity of computing power would exceed~~***which exceeds*** one hundred million dollars ($100,000,000) ~~if~~***when*** calculated using ***the*** average market ~~prices~~***price*** of cloud compute ***at the start of training*** as reasonably assessed by the developer ~~at the time of training~~.

~~(2~~***(B) An artificial intelligence model created by fine-tuning a covered model using a quantity of computing power equal to or greater than 3 times 10^25 integer or floating-point operations.***

***(2) Except as provided in subparagraph (C), on and after January 1, 2027, "covered model" means any of the following:***

***(A) An artificial intelligence model trained using a quantity of computing power determined by the Frontier Model Division pursuant to Section 11547.6 of the Government Code, the cost of which exceeds one hundred million dollars ($100,000,000) when calculated using the average market price of cloud compute at the start of training as reasonably assessed by the developer.***

***(B) An artificial intelligence model created by fine-tuning a covered model using a quantity of computing power that exceeds a threshold determined by the Frontier Model Division.***

***(C) If the Frontier Model Division does not adopt a regulation governing subparagraphs (A) and (B) by January 1, 2027, the definition of "covered model" in paragraph (1) continues to be in effect until the regulation is adopted.***

***(3)*** On and after January 1, 2026, the dollar amount in this subdivision shall be adjusted annually for inflation to the nearest one hundred dollars ($100) based on the change in the annual California Consumer Price Index for All Urban Consumers published by the Department of Industrial Relations for the most recent annual period ending on December 31 preceding the adjustment.

~~(g) "Critical harm" means a harm listed in paragraph (1) of subdivision (n).~~

~~(i) (1) "Derivative~~***(f) "Covered*** model~~" means an artificial intelligence model that is a~~ derivative ~~of another artificial intelligence model, including either~~ ***" means any*** of the following:

(~~A) A modified or~~***1) An*** unmodified copy of ~~an artificial intelligence~~***a covered*** model.

***(2) A copy of a covered model that has been subjected to post-training modifications unrelated to fine-tuning.***

***(3) (A) Before January 1, 2027, a copy of a covered model that has been fine-tuned using a quantity of computing power not exceeding 3 times 10^25 integer or floating-point operations.***

(B) ~~A combination~~***On and after January 1, 2027, a copy*** of ~~an artificial intelligence~~***a covered*** model ~~with other software.~~

~~(2) "Derivative model" does not include either of the following:~~

~~(A) An entirely independently trained artificial intelligence model.~~

~~(B) An artificial intelligence model, including one combined with other software,~~ that ~~is~~***has been*** fine-tuned using a quantity of computing power ~~greater than 25 percent of the quantity of computing power, measured in integer or floating-point operations, used to train the original model~~***not exceeding a threshold determined by the Frontier Model Division***.

~~(j) (1) "Developer" means a person that creates, owns, or otherwise has responsibility for an artificial intelligence model.~~

~~(2) "Developer" does not include a third-party machine-learning operations platform, an artificial intelligence infrastructure platform, a computing cluster, an application developer using sourced models, or an end-user of an artificial intelligence model.~~

~~(k) "Fine tuning" means the adjustment of the model weights of an artificial intelligence model after it has finished its initial training by training the model with new data.~~

~~(l) "Frontier Model Division" means the Frontier Model Division created pursuant to Section 11547.6 of the Government Code.~~

~~(m) (1) "Full shutdown" means the cessation of operation of a covered model, including all copies and derivative models, on all computers and storage devices within the custody, control, or possession of a nonderivative model developer or a person that operates a computing cluster, including any computer or storage device remotely provided by agreement.~~

~~(2) "Full shutdown" does not mean the cessation of operation of a covered model to which access was granted pursuant to a license that was not granted by the licensor on a discretionary basis and was not subject to separate negotiation between the parties.~~

~~(n) (1) "Hazardous capability" means the capability of a covered model to be used to enable any of the following harms in a way that would be significantly more difficult to cause without access to a covered model that does not qualify for a limited duty exemption:~~

***(C) If the Frontier Model Division does not adopt a regulation governing subparagraph (B) by January 1, 2027, the quantity of computing power specified in subparagraph (A) shall continue to apply until the regulation is adopted.***

***(4) A copy of a covered model that has been combined with other software.***

***(g) (1) "Critical harm" means any of the following harms caused or enabled by a covered model or covered model derivative:***

(A) The creation or use of a chemical, biological, radiological, or nuclear weapon in a manner that results in mass casualties.

(B) ~~At~~***Mass casualties or at*** least five hundred million dollars ($500,000,000) of damage ~~through~~***resulting from*** cyberattacks on critical infrastructure~~ via~~**,** ***occurring either in*** a single incident or ***over*** multiple related incidents.

(C) ~~At~~***Mass casualties or at*** least five hundred million dollars ($500,000,000) of damage ~~by~~***resulting from*** an artificial intelligence model ~~that~~ autonomously ~~engages~~***engaging*** in conduct that would ~~violate~~***constitute a serious or violent felony under*** the Penal Code if undertaken by a human with the ~~necessary~~***requisite*** mental state~~ and causes either of the following:~~**.**

~~(i) Bodily harm to another human.~~

~~(ii) The theft of, or harm to, property.~~

(D) Other grave ~~threats~~***harms*** to public safety and security that are of comparable severity to the harms described in paragraphs (A) to (C), inclusive.

~~(2) "Hazardous capability" includes a capability described in paragraph (1) even if the hazardous capability would not manifest but for fine tuning and posttraining modifications performed by third-party experts intending to demonstrate those abilities.~~

***(2) "Critical harm" does not include harms caused or enabled by information that a covered model outputs if the information is otherwise publicly accessible.***

(3) On and after January 1, 2026, the dollar amounts in this subdivision shall be adjusted annually for inflation to the nearest one hundred dollars ($100) based on the change in the annual California Consumer Price Index for All Urban Consumers published by the Department of Industrial Relations for the most recent annual period ending on December 31 preceding the adjustment.

(h) "Critical infrastructure" means assets, systems, and networks, whether physical or virtual, the incapacitation or destruction of which would have a debilitating effect on physical security, economic security, public health, or safety in the state.

~~(o) "Limited duty exemption" means an exemption, pursuant to subdivision (a) or (c) of Section 22603, with respect to a covered model that is not a derivative model, which applies if a developer can provide reasonable assurance that a covered model does not have a hazardous capability and will not come close to possessing a hazardous capability when accounting for a reasonable margin for safety and the possibility of posttraining modifications.~~

~~(p) "Machine-learning operations platform" means a solution that includes a combined offering of necessary machine-learning development capabilities, including exploratory data analysis, data preparation, model training and tuning, model review and governance, model inference and serving, model deployment and monitoring, and automated model retraining.~~

~~(q~~***(i) "Developer" means a person that performs the initial training of a covered model either by training a model using a sufficient quantity of computing power, or by fine-tuning an existing covered model using sufficient quantity of computing power pursuant to subdivision (e).***

***(j) "Fine-tuning" means adjusting the model weights of a trained covered model by exposing it to additional data.***

***(k) "Frontier Model Division" means the Frontier Model Division created pursuant to Section 11547.6 of the Government Code.***

***(l) "Full shutdown" means the cessation of operation of any of the following:***

***(1) The training of a covered model.***

***(2) A covered model.***

***(3) All covered model derivatives controlled by a developer.***

(***m***) "Model weight" means a numerical parameter ~~established through training~~ in an artificial intelligence model that ***is adjusted through training and that*** helps determine how ~~input information impacts a model's output~~***inputs are transformed into outputs***.

(~~r~~***n***) "Open-source artificial intelligence model" means an artificial intelligence model that is made freely available and ***that*** may be freely modified and redistributed.

(~~s~~***o***) "Person" means an individual, proprietorship, firm, partnership, joint venture, syndicate, business trust, company, corporation, limited liability company, association, committee, or any other nongovernmental organization or group of persons acting in concert.

(~~t) "Posttraining~~ ***p) "Post-training*** modification" means ~~the modification of~~***modifying*** the capabilities of ~~an artificial intelligence~~***a covered*** model ~~after the completion of training~~ by any means, including, but not limited to, ~~initiating additional training~~***fine-tuning***, providing the model with access to tools or data, removing safeguards against hazardous misuse or misbehavior of the model, or combining the model with, or integrating it into, other software.

(~~u~~***q***) "Reasonable assurance" does not mean full certainty or practical certainty.

(~~v~~***r***) "Safety and security protocol" means documented technical and organizational protocols that meet both of the following criteria:

(1) The protocols are used to manage the risks of developing and operating covered models across their life cycle, including risks posed by enabling or potentially enabling the creation of ~~derivative models~~***covered model derivatives***.

(2) The protocols specify that compliance with the protocols is required in order to train, operate, possess, and provide external access to the developer's covered model.

22603.

(a) Before ~~initiating training of~~***a developer initially trains*** a covered model ~~that is not a derivative model, a developer of that covered model may determine whether~~**,** the ~~covered model qualifies for a limited duty exemption.~~

~~(1) In making the determination authorized by this subdivision, a~~ developer shall ~~incorporate all applicable covered guidance.~~

~~(2) A developer may determine that a covered model qualifies for a limited duty exemption if the covered model will have lower performance on all benchmarks relevant under subdivision (f) of Section 22602 and has an equal or lesser general capability than either of the following:~~

~~(A) A noncovered model that manifestly lacks hazardous capabilities.~~

~~(B) Another model that is the subject of a limited duty exemption.~~

~~(3) Upon determining that a covered model qualifies for a limited duty exemption, the developer of the covered model shall submit to the Frontier Model Division a certification under penalty of perjury that specifies the basis for that determination.~~

~~(4) A developer that makes a good faith error regarding a limited duty exemption shall be deemed to be in compliance with this subdivision if the developer reports its error to the Frontier Model Division within 30 days of completing the training of the covered model and ceases operation of the artificial intelligence model until the developer is otherwise in compliance with subdivision (b).~~

~~(b) Before initiating training of a covered model that is not a derivative model and is not the subject of a limited duty exemption, and until that covered model is the subject of a limited duty exemption, the developer of that covered model shall~~ do all of the following:

(1) Implement administrative, technical, and physical cybersecurity protections to prevent unauthorized access to, ~~or~~ misuse ***of,*** or unsafe ~~modification~~***post-training modifications*** of~~,~~ the covered model~~, including to prevent theft, misappropriation, malicious use, or inadvertent release or escape of the~~ ***and all covered*** model ~~weights from~~***derivatives controlled by*** the ~~developer's custody,~~***developer*** that are appropriate in light of the risks associated with the covered model, including from advanced persistent threats or other sophisticated actors.

(2) Implement the capability to promptly enact a full shutdown ~~of the covered model~~.

(3) Implement ~~all covered guidance.~~

~~(4) Implement~~ a written and separate safety and security protocol that does all of the following:

(A) ~~Provides reasonable assurance that if~~***If*** a developer complies with ~~its~~***the*** safety and security protocol, ~~either of the following will apply:~~

~~(i) The~~***provides reasonable assurance that the*** developer will not produce a covered model ~~with a hazardous capability or enable the production of a~~ ***or covered model*** derivative ~~model with a hazardous capability.~~

~~(ii) The safeguards enumerated in the protocol will be sufficient to prevent~~***that poses an*** unreasonable risk of ***causing or enabling a*** critical ~~harms from the exercise of a hazardous capability in a covered model.~~***harm.***

(B) States compliance requirements in an objective manner and with sufficient detail and specificity to allow the developer or a third party to readily ascertain whether the requirements of the safety and security protocol have been followed.

(C) Identifies specific tests and test results that would be sufficient to provide reasonable assurance ~~that a covered model does not have a hazardous capability and will not come close to possessing a hazardous capability when accounting for a reasonable margin for safety and the possibility of posttraining modifications, and in addition does all of the following:~~***of the following:***

~~(i~~***(i) That a covered model does not pose an unreasonable risk of causing or enabling a critical harm.***

***(ii) That covered model derivatives do not pose an unreasonable risk of causing or enabling a critical harm.***

***(D)*** Describes in detail how the testing procedure ~~incorporates fine tuning and posttraining~~***assesses the risks associated with post-training*** modifications ~~performed by third-party experts intending to demonstrate those abilities~~.

~~(ii) Describes in detail how the testing procedure incorporates the possibility of posttraining modifications.~~

~~(iii) Describes in detail how the testing procedure incorporates the requirement for reasonable margin for safety.~~

~~(iv~~***(E)*** Describes in detail how the testing procedure addresses the possibility that a covered model can be used to make ~~posttraining~~***post-training*** modifications or create another covered model in a manner that may generate hazardous capabilities.

(~~v~~***F***) Provides sufficient detail for third parties to replicate the testing procedure.

(~~D~~***G***) Describes in detail how the developer will ~~meet requirements listed~~***fulfill their obligations*** under ~~paragraphs (1), (2), (3), and (5).~~***this chapter.***

~~(E) If applicable, describes~~***(H) Describes*** in detail how the developer intends to implement the safeguards and requirements referenced in paragraph (1~~) of subdivision (d~~).

(~~F~~***I***) Describes in detail the conditions ~~that~~***under which a developer*** would ~~require the execution of~~***enact*** a full shutdown.

(~~G~~***J***) Describes in detail the procedure by which the safety and security protocol may be modified.

~~(H) Meets other criteria stated by the Frontier Model Division in guidance to achieve the purpose of maintaining the safety of a covered model with a hazardous capability.~~

(~~5~~***(4***) Ensure that the safety and security protocol is implemented as written, including~~, at a minimum,~~ by designating senior personnel ***to be*** responsible for ensuring ~~implementation~~***compliance*** by employees and contractors working on a covered model, monitoring and reporting on implementation, and conducting audits, including through third ~~parties as appropriate~~***party auditors***.

(~~6~~***5***) Provide a copy of the safety and security protocol to the Frontier Model Division.

(~~7~~***6***) Conduct an annual review of the safety and security protocol to account for any changes to the capabilities of the covered model and industry best practices and, if necessary, make modifications to the policy.

(~~8~~***7***) If the safety and security protocol is modified, provide an updated copy to the Frontier Model Division within 10 business days.

(~~9) Refrain~~***8) Implement other reasonable measures to prevent covered models and covered model derivatives*** from ~~initiating training~~ ***posing unreasonable risks*** of ***causing or enabling critical harms.***

***(b) Before using*** a covered model ~~if there remains an unreasonable risk that an individual, or the covered model itself, may be able to use the hazardous capabilities of the~~ ***or*** covered model~~, or a~~ derivative***, or making a covered*** model ~~based on it, to cause a critical harm.~~

~~(10) Implement other measures that are reasonably necessary, including in light of applicable guidance from the Frontier Model Division, National Institute of Standards and Technology, and standard-setting organizations, to prevent the development or exercise of hazardous capabilities or to manage the risks arising from them.~~

~~(c) (1) Upon completion of the training of a~~ ***or*** covered model ~~that is not the subject of a limited duty exemption under subdivision (a) and is not a~~ derivative ~~model~~***available for commercial or public use***, the developer ~~shall perform capability testing sufficient to determine if a limited duty exemption applies with respect to the covered model pursuant to its safety and security protocol.~~

~~(2) Upon determining if a limited duty exemption applies with respect to the covered model, a developer of the covered model shall submit to the Frontier Model Division, under penalty of~~

~~perjury, a certification of compliance with the requirements of this section within 90 days and no more than 30 days after initiating the commercial, public, or widespread use of the covered model that includes both of the following:~~

~~(A) The basis for the developer's determination whether a limited duty exemption applies.~~

~~(B) The specific methodology and results of the capability testing undertaken pursuant to this subdivision.~~

~~(d) Before initiating the commercial, public, or widespread use of a covered model that is not subject to a limited duty exemption, a developer of the nonderivative version of the~~***of a*** covered model shall do all of the following:

> ~~(1)~~ ***(1) Assess whether the covered model is reasonably capable of causing a critical harm.***

***(2)*** Implement reasonable safeguards ~~and requirements, informed by~~***to prevent*** the ~~training and testing process, to do all of the following:~~

> ~~(A) Prevent an individual from being able to use the hazardous capabilities of the~~***covered*** model~~, or a derivative~~ ***and covered*** model~~, to cause~~ ***derivatives from causing*** a critical harm.

~~(B) Prevent an individual from being able to use the model to create a derivative model that is used to cause a critical harm.~~

> ~~(C~~***(3***) Ensure, to the extent reasonably possible, that the covered model's actions and ***the actions of covered model derivatives, as well as*** any ~~resulting~~ critical harms ***resulting from their actions,*** can be accurately and reliably attributed to ~~it and any user responsible~~***them.***
>
> ***(4) Beginning January 1, 2028, obtain a certificate of compliance from a third party auditor who has been accredited pursuant to 11547.6 of the Government Code.***

***(c) A developer shall not use a covered model commercially or publicly, or make a covered model or a covered model derivative available*** for ~~those actions~~***commercial or public use, if there is an unreasonable risk that the covered model or covered model derivative can cause or enable a critical harm***.

~~(2)~~ ***(d)*** A~~) Provide reasonable requirements to developers of derivative models to prevent an individual from being able to use a derivative model to cause a critical harm.~~

~~(B) If a~~ developer ~~provides access to the derivative model in a form that makes fine tuning possible, provide information to developers of that derivative model in a manner that will enable them to determine whether they have done a sufficient amount of fine tuning to meet the threshold described in subparagraph (B) of paragraph (2) of subdivision (i) of Section 22602.~~

~~(3) Refrain from initiating the commercial, public, or widespread use~~ of a covered model ~~if there remains an unreasonable risk that an individual may be able to use the hazardous capabilities of the model, or a derivative model based on it, to cause a critical harm.~~

~~(4) Implement other measures that are reasonably necessary, including in light of applicable guidance from the Frontier Model Division, National Institute of Standards and Technology, and standard-setting organizations, to prevent the development or exercise of hazardous capabilities or to manage the risks arising from them.~~

~~(e) A developer of a nonderivative covered model~~ shall ~~periodically~~***annually*** reevaluate the procedures, policies, protections, capabilities, and safeguards implemented pursuant to this section ~~in light of the growing capabilities of covered models and as is reasonably necessary to ensure that the covered model or its users cannot remove or bypass those procedures, policies, protections, capabilities, and safeguards~~.

(~~f~~***e***) (1) A developer of a ~~nonderivative~~ covered model ~~that is not the subject of a limited duty exemption~~ shall ***annually*** submit to the Frontier Model Division ~~an annual~~***a*** certification under penalty of perjury of compliance with the requirements of this section signed by the chief technology officer, or a more senior corporate officer, in a format and on a date as prescribed by the Frontier Model Division. ***This paragraph applies as long as the covered model or any covered model derivatives controlled by the developer remain in commercial or public use, or remain available for commercial or public use.***

(2) In a certification submitted pursuant to paragraph (1), a developer shall specify or provide, at a minimum, all of the following:

(A) The nature and magnitude of ~~hazardous capabilities~~***critical harms*** that the covered model ~~possesses~~ or ***covered model derivatives*** may reasonably ~~possess~~***cause or enable,*** and the outcome of ~~capability testing~~***the assessment*** required by subdivision (~~c~~***b***).

(B) An assessment of the risk that compliance with the safety and security protocol may be insufficient to prevent ~~harms from the exercise of~~ the covered ~~model's hazardous capabilities~~***model or covered model derivatives from causing critical harms***.

~~(C) Other information useful to accomplishing the purposes of this subdivision, as determined by the Frontier Model Division.~~

~~(g~~***(C) A description of the process used by the signing officer to verify compliance with the requirements of this section, including a description of the materials reviewed by the signing officer, a description of testing or other evaluation performed to support the certification, and the contact information of any third parties relied upon to validate compliance.***

***(D) Beginning January 1, 2028, a certificate of compliance from an accredited third party auditor.***

***(f)*** (1) A developer of a ~~nonderivative~~ covered model shall report each artificial intelligence safety incident affecting ~~that~~***the*** covered model ~~and any derivative version of that~~***or any*** covered model ~~within the custody, control, or possession of the~~ ***derivatives controlled by the*** developer~~,~~ ~~as described in subdivision (m) of Section 22602,~~ to the Frontier Model Division ~~in a manner prescribed by the Frontier Model Division.~~

~~(2) The report required by this subdivision shall be made not later than~~***within*** 72 hours ~~after~~***of*** the developer ~~learns that an~~ ***learning of the*** artificial intelligence safety incident ~~has occurred~~, or ***within 72 hours of*** the developer ~~learns~~***learning*** facts sufficient to establish a reasonable belief that an artificial intelligence safety incident has occurred.

~~(h) (1) (A) Reliance on an unreasonable limited duty exemption does not relieve a developer of its obligations under this section.~~

~~(B) A determination that a covered model qualifies for a limited duty exemption that results from a good faith error reported pursuant to paragraph (4) of subdivision (a) is not an unreasonable limited duty exemption.~~

~~(2) A limited duty exemption is unreasonable if the developer does not take into account reasonably foreseeable risks of harm or weaknesses in capability testing that lead to an inaccurate determination.~~

~~(3) A risk of harm or weakness in capability testing is reasonably foreseeable, if, by the time that~~***(g) A developer shall submit to the Frontier Model Division, under penalty of perjury, a certification of compliance with the requirements of this section no more than 30 days after making a covered model or covered model derivative available for commercial or public use for the first time. A developer need not submit a certification for a covered model derivative if the developer has already submitted a certification for the applicable covered model.***

***(h) In fulfilling their obligations under this chapter,*** a developer ~~releases a model, an applicable risk of harm or weakness in capability testing has already been identified by either of the following:~~

~~(A) Any other developer of a comparable or comparably powerful model through risk assessment, capability testing, or other means.~~

~~(B) By the~~***shall consider applicable guidance from the Frontier Model Division,*** National Institute of Standards and Technology, ~~the Frontier Model Division, or any independent~~***and other reputable*** standard-setting ~~organization or capability testing organization cited by either of those entities.~~***organizations.***

22604. ***(a)*** A person that operates a computing cluster shall implement ~~appropriate~~ written policies and procedures to do all of the following when a customer utilizes compute resources that would be sufficient to train a covered model:

(~~a~~***1***) Obtain a prospective customer's basic identifying information and business purpose for utilizing the computing cluster, including all of the following:

(~~1~~***A***) The identity of that prospective customer.

(~~2~~***B***) The means and source of payment, including any associated financial institution, credit card number, account number, customer identifier, transaction identifiers, or virtual currency wallet or wallet address identifier.

(~~3~~***C***) The email address and telephonic contact information used to verify a prospective customer's identity.

(~~b~~***2***) Assess whether a prospective customer intends to utilize the computing cluster to ~~deploy~~***train*** a covered model.

~~(c) Annually~~***(3) If a customer repeatedly utilizes computer resources that would be sufficient to train a covered model,*** validate the information ***initially*** collected pursuant to subdivision (a) and conduct the assessment required pursuant to subdivision (b~~).~~***) prior to each utilization.***

~~(d~~***(4) Retain a customer's Internet Protocol addresses used for access or administration and the date and time of each access or administrative action.***

***(5***) Maintain for seven years and provide to the Frontier Model Division or the Attorney General, upon request, appropriate records of actions taken under this section, including policies and procedures put into effect.

(~~e~~***6***) Implement the capability to promptly enact a full shutdown ~~in the event of an emergency.~~

~~(f) Retain a~~ ***of any resources being used to train or operate such*** customer's ~~Internet Protocol addresses used for access or administration and the date and time of each access or administrative action~~***administered models***.

***(b) A person that operates a computing cluster shall consider applicable guidance from the Frontier Model Division, National Institute of Standards and Technology, and other reputable standard-setting organizations.***

[…]

SEC. 4. Section 11547.6 is added to the Government Code, to read:

11547.6. (a) As used in this section:

(1) "~~Hazardous capability~~***Critical harm***" has the same meaning as defined in Section 22602 of the Business and Professions Code.

~~(2) "Limited duty exemption" has the same meaning as defined in Section 22602 of the Business and Professions Code.~~

~~(b~~***(b) There is hereby established the Board of Frontier Models. The board shall be housed in the Government Operations Agency and shall independent of the Department of Technology. The Governor may appoint an executive officer of the board, subject to Senate confirmation, who shall hold the office at the pleasure of the Governor. The executive officer shall be the administrative head of the board and shall exercise all duties and functions necessary to ensure that the responsibilities of the board are successfully discharged.***

***(c) Commencing January 1, 2026, the Board shall be composed of 5 members, as follows:***

***(1) A member of the open-source community, appointed by the Governor, subject to Senate confirmation.***

***(2) A member of the artificial intelligence industry, appointed by the Governor, subject to Senate confirmation.***

***(3) A member of academia, appointed by the Governor, subject to Senate confirmation.***

***(4) A member appointed by the Speaker of the Assembly.***

***(5) A member appointed by the Senate Rules Committee.***

***(d***) The Frontier Model Division is hereby created within the ~~Department of Technology~~***Government Operations Agency under the direct supervision of the Board***.

(~~c~~***e***) The Frontier Model Division shall do all of the following:

(1) Annually review certification reports received from developers pursuant to Section 22603 of the Business and Professions Code and publicly release summarized findings based on those reports.

(2) Advise the Attorney General on potential violations of this section or Chapter 22.6 (commencing with Section 22602) of Division 8 of the Business and Professions Code.

(3) (A) Issue guidance, standards, and best practices necessary to prevent unreasonable risks ~~from~~***of*** covered models ~~with hazardous capabilities~~***and covered model derivatives causing critical harms,*** including, but not limited to, more specific components of or requirements under the duties required under Section 22603 of the Business and Professions Code.

(B) Establish an ~~optional~~ accreditation process and relevant accreditation standards under which third ~~parties~~***party auditors*** may be accredited for a three-year period, which may be extended through an appropriate process, to certify adherence by developers to ~~the best practices and standards adopted pursuant to subparagraph (A).~~***their requirements under Section 22603 of the Business and Professions Code.***

(4) Publish anonymized artificial intelligence safety incident reports received from developers pursuant to Section 22603 of the Business and Professions Code.

(5) (A) Issue guidance describing the categories of artificial intelligence safety events that are likely to constitute a state of emergency within the meaning of subdivision (b) of Section 8558 and responsive actions that could be ordered by the Governor after a duly proclaimed state of emergency.

(B) The guidance issued pursuant to subparagraph (A) shall not limit, modify, or restrict the authority of the Governor in any way.

(6) Appoint and consult with an advisory committee that shall advise the Governor on when it may be necessary to proclaim a state of emergency relating to artificial intelligence and advise the Governor on what responses may be appropriate in that event.

(7) Appoint and consult with an advisory committee for open-source artificial intelligence that shall do all of the following:

(A) Issue guidelines for model evaluation for use by developers of open-source artificial intelligence models that ~~do not have hazardous capabilities~~***lack the ability to cause or enable critical harms***.

(B) Advise the ~~Frontier Model Division~~***Legislature*** on the creation and feasibility of incentives, including tax credits, that could be provided to developers of open-source artificial intelligence models that are not covered models.

(C) Advise the Frontier Model Division on future policies and legislation impacting open-source artificial intelligence development.

(8) Levy fees, including an assessed fee for the submission of a certification, in an amount sufficient to cover the reasonable costs of administering this section that do not exceed the reasonable costs of administering this section.

(9) (A) Develop and submit to the Judicial Council proposed model jury instructions for actions involving violations of Section 22603 of the Business and Professions Code that the Judicial Council may, at its discretion, adopt.

(B) In developing the model jury instructions required by subparagraph (A), the Frontier Model Division shall consider all of the following factors:

(i) The level of rigor and detail of the safety and security protocol that the developer faithfully implemented while it trained, stored, and released a covered model.

(ii) Whether and to what extent the developer's safety and security protocol was inferior, comparable, or superior, in its level of rigor and detail, to the safety and security protocols of comparable developers.

(iii) The extent and quality of the developer's safety and security protocol's prescribed safeguards, capability testing, and other precautionary measures with respect to the relevant ~~hazardous capability and related hazardous capabilities~~***risk of causing a critical harm***.

(iv) Whether and to what extent the developer and its agents complied with the developer's safety and security protocol, and to the full degree, that doing so might plausibly have avoided causing a particular harm.

(v) Whether and to what extent the developer carefully and rigorously investigated, documented, and accurately measured, insofar as reasonably possible given the state-of-the-art, relevant risks that its model might pose.

~~(10) (A) On or before July 1, 2026, issue guidance regarding both of the following:~~

***(10) (A) On or before Jan 1, 2027, and annually thereafter, issue regulations to update the definition of a "covered model" to ensure that it accurately reflects technological developments, scientific literature, and widely-accepted national and international standards and applies to artificial intelligence models that pose the greatest risk of enabling critical harms. The updated definition shall contain the following:***

(i) ~~Information relevant to determining whether~~***The initial compute threshold that*** an artificial intelligence model ~~is~~***must exceed to be considered*** a covered model, as defined in Section 22602 of the Business and Professions Code.

~~(ii) Technical thresholds and benchmarks relevant to determining whether a covered model is subject to a limited duty exemption under paragraph (2) of subdivision (a) of Section 22603 of the Business and Professions Code.~~

***(ii) The fine-tuning compute threshold that an artificial intelligence model must meet to be considered a covered model.***

(B) In developing ~~guidance~~***regulations*** pursuant to this paragraph, the Frontier Model Division shall take into account both of the following:

(i) The quantity of computing power used to train covered models that have been identified as ~~having hazardous capabilities~~***being reasonably likely to cause*** or ~~not having hazardous capabilities when accounting for~~***enable*** a ~~reasonable margin for safety~~***critical harm***.

(ii) Similar thresholds used in federal law***, guidance,*** or regulations for the management of ~~hazardous capabilities~~***models with reasonable risks of causing critical harms***.

~~(11) At least every~~***(iii) Input from stakeholders, including academics, industry, and government entities, including from the open-source community.***

***(10) Every*** 24 months after initial publication of guidance under paragraphs (3), (5), and (10), review existing guidance in consideration of technological advancements, changes to industry best practices, and information received pursuant to paragraph (1) and update its guidance to the extent appropriate.

(~~12~~***11***) On and after January 1, 2026, annually publish the inflation-adjusted dollar amounts described in paragraph (3) of subdivision (n) and paragraph (2) of subdivision (f) of Section 22602 of the Business and Professions Code.

(~~d~~***f***) There is hereby created in the General Fund the Frontier Model Division Programs Fund.

(1) All fees received by the Frontier Model Division pursuant to this section shall be deposited into the fund.

(2) All moneys in the account shall be available, only upon appropriation by the Legislature, for purposes of carrying out the provisions of this section.

[…]

***ARGUMENTS IN SUPPORT:***

The Center for AI Safety Action Fund, a co-sponsor of this bill, writes:

> If California does not act to establish a clear and sensible governance framework, the safety of its citizens could be imperiled, the nation's security could be seriously harmed, and AI's enormous potential to improve our world could be derailed. Placing sensible guardrails around serious risks that the most powerful systems might pose, while taking significant steps towards leveling the playing field for academics and startups, is the best way to ensure that CA's citizens can realize the immense benefits of this technology.

Encode Justice, a co-sponsor of this bill, writes:

> SB 1047 introduces essential safeguards for the creation of highly capable AI models, often known as "frontier AI models." These models are defined in the bill as trained using over 10^26 floating-point operations. Models of this scope would cost at least $100 million to develop and, notably, **do not yet publicly exist** but are anticipated to emerge soon as technological advancements continue. These are advanced, resource-intensive projects that have caught attention at the highest levels of government and are the focus of President Biden's Executive Order on Artificial Intelligence for their significant national security and public safety implications.

A coalition of labs, startups, and other entities including The Future Society write:

> Building on Executive Orders on artificial intelligence from Governor Newsom and the Biden Administration, and the voluntary commitments to the White House, SB 1047 sets out clear standards for developers of the largest AI models, more powerful than any model that exists today — defined as those trained with more than 1026 floating-point operations of computing power and costing more than $100 million to train. SB 1047 will help ensure that these systems are developed in a safe and secure fashion while ensuring that California remains a leader in AI development.

***ARGUMENTS IN OPPOSITION:***

California Chamber of Commerce writes on behalf of a coalition of trade associations:

> In addition to creating inconsistencies with federal regulations, the bill demands compliance with various vague and impractical, if not technically infeasible, requirements for which developers will be subject to harsh penalties, including potential criminal liability. We are concerned that the bill regulates AI technology as opposed to its high-risk applications, creates significant regulatory uncertainty and therefore high compliance costs, and poses significant liability risks to developers for failing to foresee and block any harmful use of their models by others – all of which inevitably discourages economic and technological innovation. And while recent amendments take important steps in responding to the open-source community, we remain concerned about the impact of the bill on AI research and development in California and the impact on startups. Overall, the bill still makes AI business too risky in California, particularly given the potential penalties under SB 1047.

The Chamber of Progress writes:

> It is critical that public policy foster an abundance of frontier models - open and closed alike, existing and new entrants. A plurality of models will catalyze AI application development and ultimately benefit consumers. However, SB 1047 gives the largest incumbent AI models and models built upon them ("derivative models") special treatment that will inevitably lead

to fewer upstart ("non-derivative") models. This will entrench the largest incumbent players in AI frontier model development - making them even more consequential - and undercut innovation when we should be encouraging a proliferation of approaches.

**REGISTERED SUPPORT / OPPOSITION**:

**Support**

Center for Ai Safety Action Fund (co-sponsor)
Economic Security Project Action (co-sponsor)
Encode Justice (co-sponsor)
Ae Studio
Ai Safety Student Team (HARVARD)
Apart Research
California State Council of Service Employees International Union (SEIU California)
Cambridge Boston Alignment Initiative
Causative Labs
Chapman University
Civic Ai Security Program
Denizen
Depict.ai
Elicit
Enh Alpha LLC
Far Ai, INC.
Fathom Radiant
General Agents
General Proximity
Gladstone Ai
Higher Ground Labs
Indivisible CA Statestrong
Kira Center for Ai Risks & Impacts
Latino Community Foundation
Lionheart Ventures
Loveable Labs Incorporated
MIT Ai Alignment
Ml Alignment & Theory Scholars
Momentum
Mythos Ventures
New Media Studio
Nonlinear
Normative
Panoplia Laboratories
Paper Farms
Redwood Research
Safe Ai Future
The Future Society
White Space Marketing Group

**Support If Amended**

Electronic Frontier Foundation
Oakland Privacy

**Opposition**

Acclamation Insurance Management Services
Allied Managed Care
Association of National Advertisers
California Chamber of Commerce
California Fuels and Convenience Alliance
California Land Title Association
California Manufacturers and Technology Association
Chamber of Progress
Civil Justice Association of California
Coalition of Small and Disabled Veteran Businesses
Computer & Communications Industry Association
Consumer Technology Association
Flasher Barricade Association
Insights Association
Los Angeles County Business Federation (BIZ-FED)
National Federation of Independent Business (NFIB)
Silicon Valley Leadership Group
Software and Information Industry Association
Technet

**Oppose Unless Amended**

BSA the Software Alliance
California Life Sciences

**Analysis Prepared by**: Slater Sharp / P. & C.P. / (916) 319-2200, Josh Tosney / P. & C.P. / (916) 319-2200