SENATE JUDICIARY COMMITTEE
Senator Thomas Umberg, Chair
2023-2024 Regular Session

SB 1047 (Wiener)
Version: March 20, 2024
Hearing Date: April 2, 2024
Fiscal: Yes
Urgency: No
CK

SUBJECT

Safe and Secure Innovation for Frontier Artificial Intelligence Systems Act

DIGEST

This bill requires developers of powerful artificial intelligence models and those providing the computing power to train such models to put appropriate safeguards and policies into place to prevent critical harms. The bill establishes a state entity to oversee the development of these models and calls for the creation of a public cloud computing cluster.

EXECUTIVE SUMMARY

Owing to recent advances in processing power and the rise of big data, artificial intelligence's (AI) capacity and the scope of its applications have expanded rapidly, impacting how we communicate, interact, entertain ourselves, travel, transact business, and consume media. It has been used to accelerate productivity, achieve efficiencies, liberate us from drudgery, write our college essay, help us understand and enjoy the world, upgrade the Pope's fashion, connect with each other, and live longer, fuller lives. It has also been used to constrain personal autonomy, compromise privacy and security, foment social upheaval, exacerbate inequality, spread misinformation, and subvert democracy. For good or ill, its transformative potential seems boundless.

This bill seeks to establish guardrails for the most powerful AI models to avoid the more catastrophic possibilities about which experts have raised alarms. It places a series of obligations on developers of "covered models" and providers of the cloud compute for training such models. The bill also seeks to establish a public cloud-computing cluster that facilitates equitable participation in the development and deployment of responsible AI systems. The bill is co-sponsored by the Center for AI Safety Action Fund, Economic Security California Action, and Encode Justice. It is supported by a host of tech companies and labor organizations and opposed by the Chamber of Progress and a coalition of industry associations.

PROPOSED CHANGES TO THE LAW

Existing law:

1) Establishes the California Department of Technology (CDT) within the Government Operations Agency, under the supervision of the Director of Technology (Director), also known as the State Chief Information Officer. (Gov. Code Sec. 11545(a).)

2) Provides that the duties of the Director include:
   a. advising the Governor on the strategic management and direction of the state's information technology (IT) resources;
   b. establishing and enforcing state IT strategic plans, policies, standards, and enterprise architecture, as specified;
   c. minimizing overlap, redundancy, and cost in state IT operations by promoting the efficient and effective use of information technology;
   d. providing technology direction to agency and department chief information officers to ensure the integration of statewide technology initiatives, compliance with IT policies and standards, and the promotion of the alignment and effective management of IT services;
   e. working to improve organizational maturity and capacity in the effective management of IT; and establishing performance management and improvement processes to ensure state IT systems and services are efficient and effective. (Gov. Code § 11545(b).)

3) Provides that persons are responsible, not only for the result of their willful acts, but also for an injury occasioned to another by their want of ordinary care or skill in the management of their property or person, except so far as the latter has, willfully or by want of ordinary care, brought the injury upon themselves. (Civ. Code § 1714(a).)

This bill:

1) Establishes the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.

2) Provides definitions for all relevant terms, including:
   a) "Artificial intelligence model" means a machine-based system designed to operate with varying levels of autonomy that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs that can influence physical or virtual environments.
   b) "Covered model" means an artificial intelligence model that was trained using a quantity of computing power greater than 10^26 integer or floating-point operations (FLOP) or a model that can reasonably be

expected to have similar performance capabilities as assessed by commonly used benchmarks.[1]

c) "Positive safety determination" means a determination with respect to a covered model that a developer can reasonably exclude the possibility that the model has a "hazardous capability" or may come close to possessing a hazardous capability when accounting for a reasonable margin for safety and the possibility of posttraining modifications.

d) "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:
   i. The creation or use of a chemical, biological, radiological, or nuclear weapon in a manner that results in mass casualties.
   ii. At least $500 million of damage through cyberattacks on critical infrastructure.
   iii. At least $500 million of damage by a model that autonomously engages in conduct that would be criminal if done by a human.
   iv. Other comparably severe threats to public safety and security.

e) "Hazardous capability" includes the capabilities above even if they would not manifest but for fine tuning and posttraining modifications performed by third-party experts.

f) "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.

g) "Covered guidance" means any of the following:
   i. Applicable guidance issued by the National Institute of Standards and Technology and by the Frontier Model Division.
   ii. Industry best practices, including relevant safety practices, precautions, or testing procedures undertaken by developers of comparable models, and any safety standards or best practices commonly or generally recognized by relevant experts in academia or the nonprofit sector.
   iii. Applicable safety-enhancing standards set by standards setting organizations.

h) "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.

[1] In the course of drafting previous amendments, language in the definition of covered model remained that was intended to be taken out. The author has agreed to amendments that would remove "in 2024" from Section 22602(f)(1).

3) Establishes the Frontier Model Division (FMD) within CDT and tasks them with various duties, including the following:
    a) Review annual certification reports from developers.
    b) Advise the Attorney General on potential violations of this law.
    c) Issue guidance, standards, and best practices sufficient to prevent unreasonable risks from covered models with hazardous capabilities.
    d) Establish confidential fora that are structured and facilitated in a manner that allows developers to share best risk management practices for models with hazardous capabilities in a manner consistent with state and federal antitrust laws.
    e) Establish an accreditation process and relevant accreditation standards under which third parties may be accredited to certify adherence by developers to the best practices and standards adopted.
    f) Publish anonymized artificial intelligence safety incident reports received from developers.
    g) Develop and submit to the Judicial Council proposed model jury instructions for actions brought by individuals injured by a hazardous capability of a covered model.
    h) Provide technical assistance and advice to the Legislature, upon request, with respect to artificial intelligence-related legislation.
    i) Appoint and consult with various advisory committees.

4) Requires developers of "covered models" to determine whether they can make a "positive safety determination" with respect to the model before initiating training. The developer is required to incorporate all "covered guidance." The determination shall be certified under penalty of perjury and submitted to the FMD, specifying the basis for it.

5) Requires developers of covered models unable to make such a safety determination, before training the model and until a positive safety determination can be made, to implement various safeguards. This includes:
    a) Implement administrative, technical, and physical cybersecurity protections to prevent unauthorized access to, or misuse or unsafe modification of, the covered model, that are appropriate given the associated risks.
    b) Implement the capability to enact a full shutdown or a "kill switch."
    c) Implement all covered guidance.
    d) Implement, and annually review, a written and separate safety and security protocol, as specified, that details how it is adequate to prevent critical harms and "hazardous capabilities" and specifies the testing procedures incorporated therein. The protocol shall be provided to FMD.
    e) Ensures that these protocol are implemented as written, including by designating senior personnel responsible and conducting audits, as appropriate.

f) Refrain from initiating training of 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 covered model, or a derivative model based on it, to cause a critical harm.

6) Requires a developer of a covered model without a positive safety determination, upon completion of training, to perform capability testing sufficient to determine whether a positive safety determination can be made, pursuant to its safety and security protocol.

7) Provides, that if a positive safety determination is then made, the developer shall submit to FMD a certification of compliance with the basis and specific methodology and results. This shall be done within 90 days and no more than 30 days after initiating the commercial, public, or widespread use of the covered model.

8) Provides that a developer that makes a good faith error regarding a positive safety determination shall be deemed to be in compliance if it reports its error to FMD within 30 days of completing the training and ceases operation of the model until the developer is otherwise in compliance. However, reliance on an unreasonable positive safety determination, as defined, does not relieve a developer of these obligations.

9) Requires a developer, if a positive safety determination is not made, before initiating the commercial, public, or widespread use of the covered model to implement reasonable safeguards and requirements to prevent an individual from being able to use the hazardous capabilities of the model, or a derivative model, to cause a critical harm or use the model to create a derivative model that is used to cause a critical harm. Reasonable requirements must be placed on developers of derivative models to further ensure prevention of these harms. The developer must ensure, to the extent reasonably possible, that the covered model's actions and any resulting critical harms can be accurately and reliably attributed to it and any user responsible.

10) Requires a developer to 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.

11) Establishes a continuing responsibility on developers to periodically reevaluate the procedures, policies, protections, capabilities, and safeguards implemented 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 them. For models still not subject to a positive safety determination, developers

must annually certify their compliance with these provisions to FMD. The certification must be signed by the chief technology officer, or a more senior corporate officer, as prescribed by FMD, and include details of the risks the model may pose. Safety incidents must be reported to FMD, as prescribed, and without unreasonable delay, but in no event more than 72 hours later.

12) 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.

13) Requires persons that operate computing clusters to implement appropriate written policies and procedures to do the following when a customer utilizes compute resources that would be sufficient to train a covered model:
    a) Obtain, and annually validate, a prospective customer's basic identifying information and business purpose for utilizing the computing cluster, as specified.
    b) Annually assess whether a prospective customer intends to utilize the cluster to deploy a covered model.
    c) Maintain for seven years and provide to the Frontier Model Division or the Attorney General, records of actions taken pursuant to this law.
    d) Implement the ability to promptly enact a full shutdown in the event of an emergency.

14) 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.

15) 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.

16) 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.

17) 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.

18) Tasks CDT 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.

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

COMMENTS

1. Defining AI

Reminiscent of the wave of legislation to regulate social media platforms, there are dozens of bills currently making their way through the Legislature to regulate AI. Before potentially subjecting the industry and regulators to a host of new laws, it seems imperative to first define exactly what it is we are talking about and to harmonize the definitions being used just as was done when the first comprehensive definition of "social media platform" was codified a few years ago.

The task is not necessarily a straightforward one, as pointed out by the United States Congressional Research Service:

> Defining AI is not merely an academic exercise, particularly when drafting legislation. AI research and applications are evolving rapidly. Thus, congressional consideration of whether to include a definition for AI in a bill, and if so how to define the term or related terms, necessarily include attention to the scope of the legislation and the current and future applicability of the definition. Considerations in crafting a definition for use in legislation include whether it is expansive enough not to hinder the future applicability of a law as AI develops and evolves, while being narrow enough to provide clarity on the entities the law affects. Some stakeholders, recognizing the many challenges of defining AI, have attempted to define principles that might help guide policymakers. Research suggests that differences in definitions used to identify AI-related research may contribute to significantly different analyses and outcomes regarding AI competition, investments, technology transfer, and application forecasts.[2]

[2] Laurie A. Harris, *Artificial Intelligence: Background, Selected Issues, and Policy Considerations* (May 19, 2021) Congressional Research Service, https://crsreports.congress.gov/product/pdf/R/R46795. All internet citations are current as of March 26, 2024.

There are several leading definitions of AI that can be relied on to craft a reasonably clear yet appropriately broad definition for this advancing technology. While there is no definition of AI in state law, the California Privacy Protection Agency (CPPA) has published draft regulations that provide a definition for AI:

> "Artificial intelligence" means a machine-based system that infers, from the input it receives, how to generate outputs that can influence physical or virtual environments. The artificial intelligence may do this to achieve explicit or implicit objectives. Outputs can include predictions, content, recommendations, or decisions. Different artificial intelligence varies in its levels of autonomy and adaptiveness after deployment. For example, artificial intelligence includes generative models, such as large language models, that can learn from inputs and create new outputs, such as text, images, audio, or video; and facial- or speech-recognition or -detection technology.[3]

At the federal level, the National Artificial Intelligence Act of 2020 provides the following:

> Artificial intelligence. The term "artificial intelligence" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to—
>
> (A) perceive real and virtual environments;
>
> (B) abstract such perceptions into models through analysis in an automated manner; and
>
> (C) use model inference to formulate options for information or action.[4]

Also at the federal level, the National Institute for Standards and Technology (NIST), in their Artificial Intelligence Risk Management Framework (AI RMF 1.0), defined an AI system as "an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."

At the international level, the Organization for Economic Co-operation and Development (OECD) published and then revised a definition for AI to serve as a standard across jurisdictions:

---

[3] *Draft Risk Assessment and Automated Decisionmaking Technology Regulations* (March 2024) CPPA, https://cppa.ca.gov/meetings/materials/20240308_item4_draft_risk.pdf.

[4] 15 U.S.C. § 9401. This definition was subsequently cited in President Biden's executive order on AI.

> An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.[5]

With the revision, OECD made the case for harmonization:

> Obtaining consensus on a definition for an AI system in any sector or group of experts has proven to be a complicated task. However, if governments are to legislate and regulate AI, they need a definition to act as a foundation. Given the global nature of AI, if all governments can agree on the same definition, it allows for interoperability across jurisdictions.[6]

Most recently, the European Parliament signed the EU AI Act, which defines AI as follows:

> "AI system" means a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.

In order to gain both the benefit of the expertise and compromise that went into formulating these definitions and the efficiencies that come with harmonization, the Committee, with a variety of stakeholders, including the author, have come up with the following definition to begin this process and to amend into the bill, which incorporates elements of the NIST definition into the internationally recognized formulations while eliminating unnecessary examples from within it:

> "Artificial intelligence" means an engineered or machine-based system that, for explicit or implicit objectives, infers, 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.

---

[5] *Recommendation of the Council on Artificial Intelligence* (July 11, 2023) OECD, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
[6] Stuart Russell, Karine Perset & Marko Grobelnik, *Updates to the OECD's definition of an AI system explained* (November 29, 2023) OECD, https://oecd.ai/en/wonk/ai-system-definition-update.

2. Frameworks for responsible development and accountability in AI

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.

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 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**. [emphasis added]

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.

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 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.[7]

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.

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 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

[7] *Blueprint For An AI Bill Of Rights* (October 2022) Office of Science and Technology Policy, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

> 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.[8]

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."[9]

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 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

---

[8] Future of Life Institute, *Pause Giant AI Experiments: An Open Letter* (2023) https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [emphasis in original].
[9] *Statement on AI Risk* (2023) Center for AI Safety, https://www.safe.ai/work/statement-on-ai-risk.

> 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.[10]

3. Ensuring the safe and secure development of AI

This bill takes a number of approaches in responding to concerns in AI regulation. The bill can generally be broken up into the following elements:

- Safety and security obligations imposed on developers of powerful AI models.
- Know-your-customer obligations imposed on operators of computing clusters.
- Price transparency and anti-discrimination provisions for models and clusters.
- Enforcement and whistleblower protections.
- Establishment of a Frontier Model Division within CDT.
- Creation of CalCompute.

a. Guardrails in connection with the development and deployment of large AI models

The central focus of this bill is ensuring that developers of "covered models," essentially extremely powerful AI systems, are proceeding with caution given the enormous potential for harm posed by them in the hands of malicious actors. At the core of these obligations is the "positive safety determination," which is a determination that a developer can reasonably exclude the possibility that a covered model has a hazardous capability or may come close to possessing a hazardous capability when accounting for a reasonable margin for safety and the possibility of posttraining modifications.

It is important to note the magnitude of the computing power necessary to meet the definition of covered model and, likewise, the magnitude of the hazardous capabilities that the bill seeks to prevent.

First, "covered models" are artificial intelligence models trained using a quantity of computing power greater than 10^26 integer of floating-point operations. Floating-point operations, or FLOP, is a measure of the amount of compute used in AI systems. This threshold of FLOP is currently out of reach for all but a handful of entities. Therefore the bill focuses on the most highly capable AI models, often referred to as "frontier models." Some of today's examples of frontier models are GPT-4 (OpenAI), Claude 3 (Anthropic), and Gemini Ultra (Google).

---

[10] Paul Scharre, *Future-Proofing Frontier AI Regulation* (March 2024) Center for New American Security, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Report_AI-Trends_FinalC.pdf.

The author explains this metric:

> SB 1047 uses this threshold to have developers begin testing for hazardous capabilities because it indicates that the model is larger than any model trained today, and without testing it is impossible to rule out the chance that the next generation of models will contain hazardous capabilities.
>
> The average performance of AI models is predictably related to how much computational power, measured in FLOP, was used to train them. Performance on any individual task will tend to go up with FLOP used to train, but with less predictability. The largest models trained so far are believed to have used only about 10^25 FLOP. Safety testing from AI developers and independent technical auditors suggests that, while current models do not yet possess hazardous capabilities, they display warning signs indicating that substantially larger models may possess such capabilities. As computational training power moves beyond 10^26 FLOP, therefore, it becomes prudent and reasonable to test for these capabilities before deploying them.

President Biden's executive order on AI included heightened requirements that companies developing large foundation models report on a regular basis to the federal government. This bill uses the same threshold used in connection with those requirements.

Given the rapid development and advancement of these models and the efficiency of their training, concerns have been raised that this metric will be insufficient before long. Recognizing this reality, the bill alternatively provides the following threshold for a "covered model":

> The artificial intelligence model was trained using a quantity of computing power sufficiently large that it could reasonably be expected to have similar or greater performance as an artificial intelligence model trained using a quantity of computing power greater than 10^26 integer or floating-point operations in 2024 as assessed using benchmarks commonly used to quantify the general performance of state-of-the-art foundation models.

The goal of this is ensure the models powerful enough to cause critical harms continue to be covered by the bill as metrics for determining the appropriate thresholds advance. Writing in opposition, a coalition of industry associations led by the California Chamber of Commerce raise issues with the definition:

> There is little to no certainty as to what this translates to in practice and, in any case, such thresholds will become obsolete within a year, requiring the law to change yet again. Moreover, by equating model size to risk, the definition of "covered models" is simultaneously overly broad and too narrow as smaller and/or less performant models can present much greater risks than large/higher performant ones. As a result, the bill both fails to adequately address the very real risks posed by small but malicious models and imposes significant costs on innovating performant but responsible ones.

While understanding the difficulty of creating defined parameters for ever-evolving AI systems, this definition suffers from a lack of clarity. For instance, who determines what the "benchmarks commonly used to quantify the general performance of state-of-the-art foundation models" are? The author has committed to working to increase clarity around how one can determine what models are included in the future.

Second, with regard to the magnitude of harms being targeted, the level of potential carnage required to constitute a "hazardous capability" is substantial. It must be capable of being 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:

- The creation or use of a chemical, biological, radiological, or nuclear (CBRN) weapon in a manner that results in mass casualties.
- At least $500 million of damage through cyberattacks on critical infrastructure.
- At least $500 million of damage by a model that autonomously engages in conduct that would be criminal if done by a human.

The definition also encompasses other comparably severe threats to public safety and security and those capabilities made possible by fine tuning and posttraining modifications performed by third-party experts intending to demonstrate those abilities.

Writing in support, the Future of Life Institute recommends some changes to ensure a more inclusive universe of critical harms:

> Though facilitating the creation or use of a chemical, biological, radiological, or nuclear (CBRN) weapon "in a manner that results in mass casualties" is considered a hazardous capability for a covered system, facilitating damage through cyberattacks on critical infrastructure or autonomous criminal conduct only rises to the level of a "hazardous capability" if it results in at least $500 million of damage. (Bus. & Prof. Code Sec. 22602(n)) This $500 million dollar threshold for damage is remarkably high, and would be extremely challenging to practically assess in preliminary testing. A system capable of, e.g., facilitating cyberattacks

> on critical infrastructure, should be properly assessed with appropriate steps taken to mitigate this risk, regardless of whether that potential damage would verifiably exceed $500 million. Furthermore, while the definition of hazardous capability in the context of cyberattacks and criminal conduct includes damage to property, CBRN risk only qualifies if it results in mass casualties. We recommend that the author remove the $500 million dollar threshold for qualifying as a hazardous capability in the cases of cyberattacks and criminal conduct, and include damage to property as a qualifying risk of facilitating CBRN development.

The author argues that given the unknown capabilities of future frontier models, reasonable guardrails should be in place for the largest of them. Therefore, the bill requires developers of such models to determine whether it can make a "positive safety determination" before initiating training of the model, essentially a finding that their model is not capable of being used to cause these large-scale harms. The developer would then certify, under penalty of perjury, such a determination to a division within CDT that shall be established by the bill, and discussed in more detail below, the Frontier Model Division (FMD).

The coalition in opposition argues against this approach:

> SB 1047 still makes it impossible for developers to actually make any positive safety determinations ruling out those hazardous capabilities by requiring developers to make the positive safety determination before they initiate training of the covered model. (Proposed Section 22603). Because a developer needs to test the model by training it in a controlled environment to make a positive safety determination 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.

However, the bill does not require developers to make a positive safety determination before they initiate training. The bill requires a developer to "determine whether it can make a positive safety determination" before initiating training. **In fact, it is likely that given the power and seemingly boundless potential of these models that most developers will be unable to make a positive safety determination before training it.**

The bill does not require the cessation of development and training should such a determination not be made. Rather, the bill requires developers to implement protections and protocol before initiating training of the models if they determine that a positive safety determination cannot be reached. This includes implementation of administrative, technical, and physical cybersecurity protections to prevent unauthorized access to, or misuse or unsafe modification of, the covered model, including to prevent theft, misappropriation, malicious use, or inadvertent release or

escape of the model weights from the developer's custody. Developers must also ensure there is a "kill switch," essentially the ability to cease operation of a covered model, including all copies and derivative models, on all computers and storage devices.

A detailed safety and security protocol must be put into place that includes specification of testing to be done and describes in detail how the developer will meet the various security requirements of the bill. The bill requires developers to ensure that these protocol are properly implemented as written, including by designating specific senior personnel responsible for implementation and conducting audits, as appropriate.

The bill also requires the developer to follow all "covered guidance," which is defined to include:

- Applicable guidance issued by the National Institute of Standards and Technology (NIST) and FMD.
- Industry best practices, including relevant safety practices, precautions, or testing procedures undertaken by developers of comparable models, and any safety standards or best practices commonly or generally recognized by relevant experts in academia or the nonprofit sector.
- Applicable safety-enhancing standards set by standards setting organizations.

This requirement is imposed before training, both when determining whether a positive safety determination can be made, and if not, thereafter. The goal is to ensure that AI developers incorporate safety practices that may be generally acknowledged as necessary for managing hazardous capabilities of covered models. However, the coalition in opposition argues the language is unworkable. They write:

> SB 1047 also requires that a developer "incorporate all covered guidance" before making a positive safety determination. However, industry and others are still trying to ascertain how to define what constitutes a high-capable, foundational model and it is unclear what will qualify as "industry best practices" or "standards setting organizations" for the purpose of incorporating all covered guidance. Such regulatory uncertainty will inevitably discourage economic and technological innovation. It would make far more sense to let the U.S. AI Safety Institute to complete its work first, after which safety and security protocols tied to those safety standards could be considered.

The open-ended definition of covered guidance is intentional as the author points out that it is not yet possible to specify all appropriate guidelines in full detail, or the specific organizations that will provide them, as the field will be adapting over time. However, greater clarity around the obligations imposed on developers may be necessary as the current language is arguably overly vague. For instance, while reasonableness is a common standard in tort law, its appearance in so many of the

definitions and standards in the bill arguably causes concern regarding the interpretability of the bill. Accordingly, the author has committed to continuing to work on finding the right balance on this front. The goal is to have widely accepted safety and security standards put into place and to ensure they are implemented properly. The language in the bill must ensure there are concrete benchmarks and standards both to ensure developers have clear direction but also for regulators and public prosecutors to have the ability to effectively hold them to those standards.

After completing training of a covered model that is not subject to a positive safety determination, developers are again required to perform specified testing pursuant to its safety and security protocol to attempt to make such a determination. If such a determination is made, the developer must submit certification to FMD within 90 days but no later than 30 days after initiating the commercial, public, or widespread use of the covered model. The author may wish to harmonize the certification process at all points in the development process with that provided for positive safety determinations made at the pre-training phase.

The bill does not prohibit initiating such widespread use of a model where a positive safety determination has not been made. Rather it requires the developer to implement reasonable safeguards and other security measures to prevent an individual from being able to use the hazardous capabilities of the model to cause a critical harm or to use the model to create a derivative model to cause a critical harm. The developer is required to 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. Developers must ensure, to the extent reasonably possible, that the covered model's actions and any resulting critical harms can be accurately and reliably attributed to it and any user responsible for those actions.

However, the developer is prohibited 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.

For those models still not subject to a positive safety determination, developers are required to annually certify compliance with the safety and security requirements of the bill to FMD.

Again, these provisions are generally aligned with the focus in President Biden's executive order with respect to the development of models of this size. The EO orders the Secretary of Commerce to require developers of these models to report, on an ongoing basis, to the federal government regarding the training and development of these models and the cybersecurity protections taken; the ownership and possession of the model weights; and the results of performance testing. In fact, many companies

have already voluntarily committed to follow specified guidelines that track with the bill's security and safety obligations. As described in the White House fact sheet:

> President Biden [convened] seven leading AI companies at the White House [] – Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI – to announce that the Biden-Harris Administration has secured voluntary commitments from these companies to help move toward safe, secure, and transparent development of AI technology.
>
> Companies that are developing these emerging technologies have a responsibility to ensure their products are safe. To make the most of AI's potential, the Biden-Harris Administration is encouraging this industry to uphold the highest standards to ensure that innovation doesn't come at the expense of Americans' rights and safety.
>
> These commitments, which the companies have chosen to undertake immediately, underscore three principles that must be fundamental to the future of AI – safety, security, and trust – and mark a critical step toward developing responsible AI. As the pace of innovation continues to accelerate, the Biden-Harris Administration will continue to remind these companies of their responsibilities and take decisive action to keep Americans safe.[11]

Some of the relevant commitments are:

> Ensuring Products are Safe Before Introducing Them to the Public
>
> - The companies commit to internal and external security testing of their AI systems before their release. This testing, which will be carried out in part by independent experts, guards against some of the most significant sources of AI risks, such as biosecurity and cybersecurity, as well as its broader societal effects.
> - The companies commit to sharing information across the industry and with governments, civil society, and academia on managing AI risks. This includes best practices for safety, information on attempts to circumvent safeguards, and technical collaboration.

[11] *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI* (July 21, 2023) The White House, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

> Building Systems that Put Security First
>
> - The companies commit to investing in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights. These model weights are the most essential part of an AI system, and the companies agree that it is vital that the model weights be released only when intended and when security risks are considered.
> - The companies commit to facilitating third-party discovery and reporting of vulnerabilities in their AI systems. Some issues may persist even after an AI system is released and a robust reporting mechanism enables them to be found and fixed quickly.

Therefore, the very companies that would be regulated by this bill committed to implementing similar safeguards. In testimony before the United States Senate Committee on the Judiciary's Subcommittee on Privacy, Technology, and the Law, Sam Altman, CEO of OpenAI made clear how important such safeguards are, stressing that "it is vital that AI companies–especially those working on the most powerful models–adhere to an appropriate set of safety requirements, including internal and external testing prior to release and publication of evaluation results."[12] He detailed some of the rigorous measures undertaken before initiating widespread use of such models:

> Prior to releasing each new version of our models, OpenAI conducts extensive testing, engages external experts for feedback, improves the model's behavior with techniques like reinforcement learning from human feedback (RLHF), and implements safety and monitoring systems.
>
> The release of our latest model, GPT-4, provides an illustrative example. After we developed GPT-4, we spent more than 6 months evaluating, testing, and improving the system before making it publicly available.
>
> In addition to our own evaluations, we engaged with external AI safety experts in a process known as "red teaming," through which they helped identify potential concerns with GPT-4 in areas including the generation of inaccurate information (known as "hallucinations"), hateful content, disinformation, and information related to the proliferation of conventional and unconventional weapons.
>
> This process helped us to better understand potential usage risks and ways to address those risks. In each of these areas, we developed

[12] Sam Altman, *Written Testimony before the U.S. Senate Committee on the Judiciary*, https://www.judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Bio%20&%20Testimony%20-%20Altman.pdf.

> mitigations to increase safety in significant ways. Some of our work involved making adjustments to the data used to train the model, during what is called the pre-training stage. Other interventions took place after initial training of the model.

However, opposition argues that the certification requirements in this bill are "unaccountably broad and hopelessly unworkable." Chamber of Progress writes:

> Unfortunately, SB 1047 forces model developers to engage in speculative fiction about imagined threats of machines run amok, computer models spun out of control, and other nightmare scenarios for which there is no basis in reality.
>
> Instead, SB 1047 forces developers operating in the real world to proactively mitigate against every conceivable harm - and many inconceivable ones - not just by the model itself, but subsequent third parties who make use of the model.

As stated, the author has committed to continuing to work to hone the specific requirements laid out in the bill. The author states the need for these provisions:

> With the political prospects for joint action between the White House and Congress in doubt, California has an indispensable role to play in ensuring that we develop this extremely powerful technology with basic guardrails, in order to allow society to experience its massive potential benefits. Clarifying that developers of the largest and most powerful AI models must take basic precautions against serious risks to public safety and national security is the clear way forward. Establishing these duties, in line with industry-leading best practices, will not address all of AI's risks and harms, but it is a necessary step forward. And no state is better positioned to take on this challenge than California.

The concerns about workability and clarity are important to engage with, and honing the precise protocols that should be put in place are critical as this bill moves through the legislative process. Establishing the exact protocols and policies that should be put into place for such a rapidly-advancing technology is inherently difficult. However, the author may wish to create more clarity in the defined terms and the specified guidance that should be followed. Just as important, the author may wish to bolster the role of FMD and the Attorney General, and possibly other entities, to ensure there is proper oversight of this development and to confirm that the appropriate and necessary safeguards are being put into place, perhaps through authorizing or mandating government or accredited third-party auditing. Again the Future of Life Institute provides suggestions for more robust oversight:

> Before initiating training of a covered model, the bill in print requires a developer to "ensure that the safety and security protocol is implemented as written" and recommends doing so by, among other things, "conducting, audits, including through third parties as appropriate." (Bus. & Prof. Code Sec. 22603(b)(5)) Third party auditing is essential to ensuring that ulterior motives do not compromise the integrity of safety protocol assessment, especially when it comes to systems potentially capable of critical harm. Put simply, Big Tech companies should not be grading their own homework with respect to reasonable safety protocols. To ensure independent assessment and accountability to the spirit of the bill, we recommend that the author consider amending the bill to: 1) authorize the Frontier Model Division to audit compliance with the bill's requirements as appropriate, or 2) require independent audits by third parties certified by the Frontier Model Division to perform them to the standards laid out by the bill and other guidance.

Given the level of complexity and the technological sophistication that will be necessary to adequately oversee this development, reliance on such independent audits may be critical as well as a focus on building up the resources of the FMD.

It should be noted that Chamber of Progress' assertions that the threats of critical harm from advanced AI systems are "imagined" and have "no basis in reality" stands in direct contrast to the very clear public proclamations from industry leaders and leading academics in the world of AI. As discussed more thoroughly above, "mitigating the risk of extinction from AI should be a global priority." Geoffrey Hinton, the godfather of AI referenced above, says about the bill:

> Forty years ago when I was training the first version of the AI algorithms behind tools like ChatGPT, no one - including myself - would have predicted how far AI would progress. Powerful AI systems bring incredible promise, but the risks are also very real and should be taken extremely seriously.
>
> SB 1047 takes a very sensible approach to balance those concerns. I am still passionate about the potential for AI to save lives through improvements in science and medicine, but it's critical that we have legislation with real teeth to address the risks. California is a natural place for that to start, as it is the place this technology has taken off.

The other godfather, Dr. Bengio, shares this perspective:

> AI systems beyond a certain level of capability can pose meaningful risks to democracies and public safety. Therefore, they should be properly tested and subject to appropriate safety measures. This bill offers a

practical approach to accomplishing this, and is a major step toward the requirements that I've recommended to legislators.

b. Operators of computing clusters: know your customers

The bill also imposes security and safety obligations on those providing the cloud compute to train these larger models. "Computing cluster" is defined as 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.

The bill requires a person that operates a computing cluster to obtain a prospective customer's basic identifying and contact information and the business purpose for utilizing the computing cluster. This includes 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. The operator of the computing cluster must also obtain the Internet Protocol addresses used for access or administration and the date and time of each access or administrative action.

The operator must also, at least annually, assess whether the customer intends to use the computing cluster to deploy a covered model and to validate the above information. Records of actions taken and policies put into place to satisfy these provisions must be maintained for seven years and provided to FMD or the Attorney General upon request. Operators must also implement the capability to promptly enact a full shutdown in the event of an emergency.

All of these duties are triggered only when a customer utilizes enough compute resources that would be sufficient to train the massive models covered by the bill. The author explains that this section of the bill is requiring "companies that provide cloud compute for frontier model training to institute 'know your customer' policies to help prevent the dangerous misuse of AI systems by malicious actors and geopolitical adversaries."

c. Enforcement

The Attorney General is authorized to bring a civil action against those reasonably believed to be in violation of these provisions. The Attorney General may seek preventative relief ranging from a temporary injunction to deletion of the covered model and the weights utilized. However, any such relief is only available in response to harm or an imminent risk or threat to public safety.

Courts are authorized to impose civil penalties in an amount not exceeding 10 percent of the cost, excluding labor cost, to develop the covered model for a first violation and

in an amount not exceeding 30 percent of the cost, excluding labor cost, to develop the covered model for any subsequent violation. In the apportionment of penalties assessed pursuant to this section, defendants shall be jointly and severally liable.
Anticipating potential corporate shell games, the bill provides that the court shall disregard corporate formalities and impose joint and several liability on affiliated entities for purposes of effectuating the intent of the bill if the court concludes that both of the following are true:

- Steps were taken in the development of the corporate structure among affiliated entities to purposely and unreasonably limit or avoid liability.
- The corporate structure of the developer or affiliated entities would frustrate recovery of penalties or injunctive relief under this section.

To ensure violations are brought to light by those closest to the actual development of these models, the bill provides whistleblower protections prohibiting developers from preventing employees from disclosing information about compliance issues to the Attorney General or retaliating against employees when they do so.

Notwithstanding the above, the bill provides that a developer that makes a good faith error regarding such a determination is still considered in compliance so long as the developer reports its error to FMD within 30 days of completing the training of the covered model and ceases operation of it until the developer is otherwise in compliance. However, in another section, the bill states that reliance on an *unreasonable* positive safety determination does not relieve a developer of its obligations under this section. A positive safety determination 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. They are reasonably foreseeable, if, by the time that a developer releases a model, an applicable risk of harm or weakness in capability testing has already been identified by another developer of a comparable model or the United States Artificial Intelligence Safety Institute, the FMD, or any independent standard-setting organization or capability-testing organization cited by either of those entities.

d. Establishment of the FMD

To serve as the central governmental entity overseeing the development of frontier AI models in California, the bill establishes the Frontier Model Division within CDT. As seen from the discussion above, the FMD will receive and maintain the reports, certifications, policies and protocol, and other information required to be submitted by developers. The bill also lays out a series of other duties, including:

- Advise the Attorney General on potential violations.

- Issue guidance, standards, and best practices sufficient to prevent unreasonable risks from covered models with hazardous capabilities including, but not limited to, more specific requirements on the duties required of developers.
- Establish an accreditation process and relevant accreditation standards under which third parties may be accredited to certify adherence by developers to the best practices and standards.
- Publish anonymized artificial intelligence safety incident reports received from developers pursuant to Section 22603 of the Business and Professions Code.
- Establish confidential fora that are structured and facilitated in a manner that allows developers to share best risk management practices for models with hazardous capabilities in a manner consistent with state and federal antitrust laws.
- 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.
- Appoint and consult with an advisory committee for open-source artificial intelligence that, among other things, will advise FMD on future policies and legislation impacting open-source artificial intelligence development.
- Provide technical assistance and advice to the Legislature, upon request, with respect to artificial intelligence-related legislation.
- Develop and submit to the Judicial Council proposed model jury instructions for actions brought by individuals injured by a hazardous capability of a covered model.

e. Responding to market dynamics

As industry races toward developing larger, more powerful AI models and seeks to commodify the seemingly infinite applications of AI, concerns are growing about the diminishing role that researchers, academic institutions, and more public-focused entities are playing in the development of AI. As reported by the Washington Post:

> As such tech behemoths as Meta, Google and Microsoft funnel billions of dollars into AI, a massive resources gap is building with even the country's richest universities. Meta aims to procure 350,000 of the specialized computer chips — called GPUs — that are essential to run the gargantuan calculations needed for AI models. In contrast, Stanford's Natural Language Processing Group has 68 GPUs for all of its work.
>
> To obtain the expensive computing power and data required to research AI systems, scholars frequently partner with tech employees. Meanwhile, tech firms' eye-popping salaries are draining academia of star talent.

> Big tech companies now dominate breakthroughs in the field. In 2022, the tech industry created 32 significant machine learning models, while academics produced three, a significant reversal from 2014, when the majority of AI breakthroughs originated in universities, according to a Stanford report.
>
> Researchers say this lopsided power dynamic is shaping the field in subtle ways, pushing AI scholars to tailor their research for commercial use. Last month, Meta CEO Mark Zuckerberg announced that the company's independent AI research lab would move closer to its product team, ensuring "some level of alignment" between the groups, he said.
>
> "The public sector is now significantly lagging in resources and talent compared to that of industry," said [Fei-Fei] Li, a former Google employee and the co-director of the Stanford Institute for Human-Centered AI. "This will have profound consequences because industry is focused on developing technology that is profit-driven, whereas public-sector AI goals are focused on creating public goods."
> …
> As Silicon Valley races to build chatbots and image generators, it is drawing would-be computer science professors with high salaries and the chance to work on interesting AI problems. Nearly 70 percent of people with PhDs in AI end up in private industry compared with 21 percent of graduates two decades ago, according to a 2023 report.[13]

The bill seeks to address this by first creating "CalCompute," a public cloud computing cluster created by CDT. The focus of the project will be conducting research into the safe and secure deployment of large-scale artificial intelligence models and fostering equitable innovation that includes:

- A fully owned and hosted cloud platform.
- Necessary human expertise to operate and maintain the platform.
- Necessary human expertise to support, train, and facilitate use of CalCompute.

CDT is specifically directed to commission an array of consultants and direct them to incorporate specified considerations into their plan. Two important considerations are:

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

---

[13] Naomi Nix, Cat Zakrzewski & Gerrit De Vynck, *Silicon Valley is pricing academics out of AI research* (March 10, 2024) The Washington Post, https://www.washingtonpost.com/technology/2024/03/10/big-tech-companies-ai-research/.

- 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.

CDT is required to report annually to the Legislature. Fealty to these, and the other stated, considerations is critical for the project to meet its lofty goals. The author may wish to further fortify, including through clear transparency and accountability measures, this aspect of the program to ensure that these resources and capabilities are truly working toward the public benefit and avoid cooption by the larger industry. The Future of Life Institute provides similar recommendation:

> Gov. Code Sec. 11547.7(e) of the bill in print authorizes the California Department of Technology to "receive private donations, grants, and local funds, in addition to allocated funding in the annual budget, to effectuate" the creation of its public cloud computing infrastructure. We recognize that budgetary constraints may limit the availability of public funds to realize the promising vision of this project, but we caution that dependence on private funds, which will likely come from the very Big Tech companies developing risky AI systems subject to the bill's requirements, can implicitly facilitate regulatory capture. Without proper constraints, this ambitious and impactful state project, if reliant on private donations for its subsistence, could succumb to influence from the profit-driven interests of industry leaders with the most disposable resources, and associated regulatory efforts in the future may similarly be beholden to these interests. We recommend that the author consider clearly defining limitations on the identification and input of private donors in this process, and explicitly prioritize the use of public funds over private funds where possible.

The Chamber of Progress, writing in opposition to the bill, commends the inclusion of CalCompute.

In addition to creating this public resource, the bill addresses the issues raised above by requiring developers of covered models that provide commercial access to it to provide a transparent, uniform, publicly available price schedule for the purchase of access to that covered model at a given level of quality and quantity subject to the developer's terms of service and prohibits them from engaging in unlawful discrimination or noncompetitive activity in determining price or access. The same obligations are imposed on operators of computing clusters with regard to access to their computing clusters. The author explains this section of the bill: "[I]n order to ensure that smaller startup developers have equal opportunities to larger players, SB 1047 requires cloud-computing companies and frontier model developers to provide transparent pricing and avoid price discrimination."

Economic Security California Action, a co-sponsor of the bill, explains the need for these provisions:

> The AI sector is seemingly very dynamic, with weekly headlines detailing major players buying and selling, investing in and spinning out different enterprises. However, all that movement belies the fact that AI is inarguably among America's most concentrated markets, with three large companies – Amazon, Google, and Microsoft – controlling access to two of the most expensive resources needed for AI: cloud computing power and large models. These same companies also own both upstream and downstream business interests, allowing for preferential pricing or deals that benefit some competitors over others. As a result, small entities and start-ups – those not owned by the big players – are at a significant disadvantage. To foster a competitive and innovative AI sector, SB 1047 requires cloud-computing companies and frontier model developers to provide fair and transparent pricing and disallows price discrimination.

4. Stakeholder positions

According to the author:

> 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.

The Center for AI Safety, a co-sponsor of the bill, explains the need for legislative action:

> Universities, startups, and technology companies across California are harnessing AI to enhance scientific research, spur new economic opportunities, and boost human creativity. With continued responsible

> development, AI has vast potential to improve Californians' lives in areas from healthcare to sustainability.
>
> However, leading AI researchers have warned that failure to take appropriate precautions could have severe consequences for public safety and security as developers produce increasingly powerful AI models. In a 2023 letter organized by the Center for AI Safety, hundreds of AI experts, academics, and industry leaders stated that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." . . .
>
> Just within this past week, SEC Chair Gary Gensler warned that unchecked AI systems could lead to a future financial meltdown, a group of eminent scientists in the US and China stressed that countries must set "red lines" around particular kinds of risky AI development, and a new spate of impressive demonstrations were released showcasing powerful AI agents that are capable of writing code and taking hundreds of independent actions without human oversight. There are warning signs that the next generation of models may have novel dangerous capabilities. Independent researchers at METR (a Berkeley-based non-profit formerly known as ARC Evals), who performed early safety testing on some of OpenAI and Anthropic's latest model release[s], has explicitly said that "for systems more capable than Claude and GPT-4, we are now at the point where we need to check carefully that new models do not have sufficient capabilities to replicate autonomously or cause catastrophic harm – it's no longer obvious that they won't be able to."

Encode Justice, another co-sponsor of the 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.
>
> SB 1047's Key Safeguards for Frontier AI:
>
> - Pre-Deployment: Developers need to rigorously test these AI giants for any risks or unexpected behaviors.

- Cybersecurity: To block hackers and prevent misuse, tight security measures are a must to protect these models from being compromised.
- Limiting Harmful Uses: If testing reveals any potentially dangerous abilities within these models, developers must restrict these capabilities to prevent misuse.
- Ongoing Monitoring: After these models go live, their behavior must be continuously monitored. Developers need to be ready to address new risks as they emerge, including shutting down the model if necessary.

SB 1047 focuses on the developers of the most advanced models, promoting best practices already in use by leading companies. This law doesn't burden smaller AI startups or the broader business and academic community using AI tools.

Chamber of Progress argues in opposition:

> 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.
>
> The disparate treatment is unaccountable and is not rationalized in the legislative text. Indeed it cannot be rationalized. Any regulation should apply fairly across approaches. SB 1047 would have the effect of freezing model innovation in its place.
>
> Developers build applications on top of models; competition at the model level will mean increased innovation at a lower cost. This in turn will promote the equitable diffusion of AI's benefits across California.

In response to these and other concerns regarding the bill's impact on innovation, the author states:

> SB 1047's requirements only apply to a very small handful of frontier model AI developers that are training the largest, most capital intensive models - today costing in excess of $100 million dollars. The vast majority of AI startups, and all AI application and use-case developers, would have

> zero new duties under SB 1047. And SB 1047's duties are modeled on responsible best practices pioneered by leading developers in the industry itself.

The coalition of industry associations, including the Association of National Advertisers, writes in opposition:

> We cannot overemphasize the importance of ensuring consistency in the AI regulatory landscape, nationally, and the need to follow federal guidance on certain issues that transcend national borders. Relevant to this bill, in October 2023, the White House issued an Executive Order (EO) that requires companies that are developing any foundation model that poses a serious risk to national security, national economic security, or national public health and safety to notify the federal government when training the model and share the results of all red-team safety tests to ensure that AI systems are safe, secur[e] and trustworthy before companies make them public.
>
> While we appreciate that in some respects, SB 1047 appears in line with the goals of the federal government and the White House's EO, the National Institute of Standards and Technology (NIST) is already working with other agencies at the federal level to establish testing and safety guidelines for large models. If enacted, SB 1047 would likely result in confusion about the correct standards to apply and place additional burdens on AI developers without commensurate gains in safety, especially as it fails to align with regulations nationally and introduces novel concepts and standards including around the assessment of what is a "hazardous capability". Indeed, given the definition of "covered models" under this bill which also scopes in any fine-tuning by downstream customers and users, SB 1047 is more far-reaching than anything seen to date in voluntary commitments, federal guidance, or the laws of any other countries.
>
> Ultimately, enacting a patchwork of inconsistent AI regulations that go into as much detail as SB 1047, will further fracture the U.S. regulatory landscape.

SUPPORT

Center for AI Safety Action Fund (sponsor)
Economic Security Project Action (sponsor)
Encode Justice (sponsor)
AE Studio
AI Safety Student Team (Harvard)
Apart Research

Cambridge Boston Alignment Initiative
Causative Labs
Civic AI Security Program
Denizen
Depict.ai
District Council of Iron Workers of the State of California and Vicinity
Elicit
Enh Alpha LLC
Far AI, INC.
Fathom Radiant
Foresight Institute
Forhumanity
Future of Life Institute
General Agents
General Proximity
Gladstone AI
Higher Ground Labs
Imbue
Indivisible CA Statestrong
Kira Center for Ai Risks & Impacts
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

OPPOSITION

Association of National Advertisers
California Chamber of Commerce
California Manufacturers and Technology Association
Chamber of Progress
Civil Justice Association of California
Computer and Communications Industry Association
Insights Association

Software and Information Industry Association
Technet

RELATED LEGISLATION

Pending Legislation:

SB 313 (Dodd, 2023) establishes the Office of Artificial Intelligence. It requires state agencies to disclose when they are using generative artificial intelligence to communicate with a person and to provide them an option to speak with a natural person at the agency. SB 313 was held on suspense in the Senate Appropriations Committee.

SB 398 (Wahab, 2023) establishes the Artificial Intelligence for California Research Act, which requires CDT to develop and implement a comprehensive research plan to study the feasibility of using advanced technology to improve state and local government services. SB 398 was never heard in the Senate Governmental Organization Committee.

SB 892 (Padilla, 2024) requires CDT to establish safety, privacy, and nondiscrimination standards relating to artificial intelligence services, as defined. Commencing August 1, 2025, the bill prohibits a contract for AI services, as defined, from being entered into by the state unless the provider meets those standards. CDT is required to report to the Legislature regarding the standards it establishes. SB 892 is currently in the Senate Governmental Organization Committee.

SB 893 (Padilla, 2024) requires the Government Operations Agency, the Governor's Office of Business and Economic Development, and CDT to collaborate to establish the California Artificial Intelligence Research Hub in the Government Operations Agency, as prescribed. SB 893 requires the hub to serve as a centralized entity to facilitate collaboration between government agencies, academic institutions, and private sector partners to advance AI research and development that seeks to harness the technology's full potential for public benefit while safeguarding privacy, advancing security, and addressing risks and potential harms to society, as prescribed. SB 893 is currently in the Senate Governmental Organization Committee.

SB 896 (Dodd, 2024) requires the Government Operations Agency, the Department of Technology, and the Office of Data and Innovation to produce a State of California Benefits and Risk of Generative Artificial Intelligence Report that includes certain items, including an examination of the most significant, potentially beneficial uses for deployment of generative artificial intelligence tools by the state, and would require those entities to update the report, as prescribed. The bill would require, as often as is deemed appropriate by the Director of Emergency Services, the California Cybersecurity Integration Center, and the State Threat Assessment Center, those entities to perform a joint risk analysis of potential threats posed by the use of generative

artificial intelligence to California's critical energy infrastructure, including those that could lead to mass casualty events and environmental emergencies.
The bill also requires a state agency or department that utilizes generative artificial intelligence to directly communicate with a person, either through an online interface or telephonically, to clearly and in a conspicuous manner identify to that person that the person's interaction with the state agency or department is being communicated through artificial intelligence. The also requires an automated decisionmaking system, as defined, utilized by a state agency or department to be evaluated for risk potential before adoption, as specified. SB 896 is currently in the Senate Governmental Organization Committee.

SB 942 (Becker, 2024) establishes the California AI Transparency Act, which, among other things, requires a covered provider, as defined, to create an AI detection tool by which a person can query the covered provider as to the extent to which text, image, video, audio, or multimedia content was created, in whole or in part, by a generative AI system, as defined, provided by the covered provider that meets certain criteria. Covered providers are required to include in AI-generated content a visible disclosure that, among other things, includes a clear and conspicuous notice, that identifies the content as generated by AI. SB 942 requires a covered provider to register with CDT and provide them a URL to any AI detection tool it has created. SB 942 is currently in this Committee.

SCR 17 (Dodd, 2023) affirms the California Legislature's commitment to President Biden's vision for a safe AI and the principles outlined in the "Blueprint for an AI Bill of Rights" and expresses the Legislature's commitment to examining and implementing those principles in its legislation and policies related to the use and deployment of automated systems. SCR 17 is currently in the Assembly Privacy and Consumer Protection Committee.

AB 331 (Bauer-Kahan, 2023) prohibits "algorithmic discrimination," that is, use of an automated decision tool to contribute to unjustified differential treatment or outcomes that may have a significant effect on a person's life. It requires any deployer of an automated decision tool to perform an impact assessment for those tools and to notify any natural person that is the subject of the consequential decision that an automated decision tool is being used to make, or be a controlling factor in making, the consequential decision. AB 331 was held in the Senate Appropriations Committee.

AB 2013 (Irwin, 2024) requires, on or before January 1, 2026, a developer, as defined, of an AI system or service to post on the developer's website documentation regarding the data used to train the AI system or service, as specified. AB 2013 is currently in the Assembly Privacy and Consumer Protection Committee.

AB 2930 (Bauer-Kahan, 2024) requires, among other things, a deployer and a developer of an automated decision tool to, on or before January 1, 2026, and annually thereafter,

perform an impact assessment for any automated decision tool the deployer uses that includes, among other things, a statement of the purpose of the automated decision tool and its intended benefits, uses, and deployment contexts. The assessments must be provided to the Civil Rights Department within 7 days of a request. AB 2930 requires a deployer to, at or before the time an automated decision tool is used to make a consequential decision, notify any natural person that is the subject of the consequential decision that an automated decision tool is being used to make, or be a controlling factor in making, the consequential decision and to provide that person with, among other things, a statement of the purpose of the automated decision tool.
AB 2930 is currently in the Assembly Privacy and Consumer Protection Committee.

Prior Legislation:

AB 302 (Ward, Ch. 800, Stats. 2023) requires the California Department of Technology (CDT), in coordination with other interagency bodies, to conduct a comprehensive inventory of all high-risk automated decision systems (ADS) used by state agencies on or before September 1, 2024, and report the findings to the Legislature by January 1, 2025, and annually thereafter, as specified.

AB 13 (Chau, 2021) would have established the Automated Decision Systems Accountability Act, which, in the context of the State's procurement policies, promotes oversight over ADS that pose a high risk of adverse impacts on individual rights. The bill was eventually gutted and amended to address a different topic.

SB 444 (Umberg, 2019) would have requested the Regents of the University of California (UC) to enact a resolution authorizing the law schools at UC Berkeley and UC Irvine to participate in a pilot project to develop AI or machine-learning solutions to address access to justice issues faced by self-representing litigants in their respective courts. The bill died in the Assembly Higher Education Committee.

AB 1576 (Calderon, 2019) would have required the Secretary of Government Operations to appoint participants to an AI working group to evaluate the uses, risks, benefits, and legal implications associated with the development and deployment of AI by California-based businesses. The bill was held on the Senate Appropriations Committee suspense file.

SJR 6 (Chang, Res. Ch. 112, Stats. 2019) urged the President and the Congress of the United States to develop a comprehensive AI Advisory Committee and to adopt a comprehensive AI policy.

ACR 215 (Kiley, Resolution Ch. 206, Stats. 2018) expressed the Legislature's support for a set of principles for the governance of AI known as the 23 Asilomar AI Principles.

****