

[SB 1047](#) Policy Memorandum

CS 182

Mehran Sahahmi, Daniel Ho

Matthew Torre, Remington Graham, Kyrán Romero, Shuvi Jha

1. Theory of Impact

SB 1047, the "Safe and Secure Innovation for Frontier Artificial Intelligence Models Act," seeks to mitigate harms of developing and applying large-scale AI technologies. SB 1047 disrupts the cycle of AI regulation lagging behind innovation by creating pre-qualification standards that steer models to meet critical risk assessment and safety standards *before* their deployment. By requiring compliance with such protective measures by computer infrastructure developers and suppliers, the policy puts society in an advantageous position by enabling proactive management of the risk with AI in contrast to the need for reactive actions afterwards.

Left unregulated, hazardous capabilities of models can lead to far-reaching societal implications—ranging from influencing democratic elections by AI-created deepfakes, to causing real physical harm by enabling weapon creation. With the creation of risk assessment standards for high-risk AI models, SB 1047 seeks to counteract such threats, ensuring that the advancement in AI technology is in line with public safety and ethical principles.

This proposal symbolizes a shift from passive compliance legislation to models of active regulation. Prior to this, many large AI companies/institutions relied on self-regulation to implement safeguards for their models. SB 1047 addresses any lack of accountability in self-regulation by providing an enforceable legal framework for these systems to abide by. This shift in framework is very similar to the European Union's AI Act that works to classify AI systems by the levels of risk they carry, implementing stricter regulations for models that have higher perceived risks. This policy makes California a leading contributor to the development of AI policy— providing safeguards for humanity while allowing for innovation to persist.

Looking beyond just short term risk mitigation, this bill sets a precedent for ethical and responsible use of AI. Future regulations aiming to balance technological innovation and public

safety can use this proposal as a model to expand and influence the reach of these proactive measures. In discussing the many harms and risks associated with large-scale AI models, SB 1047 provides solutions and sets the stage for a future of safe, sustainable, ethical AI developments.

2. Analysis of Main Components

This section is an assessment of the broad strokes of SB 1047. The bill is nuanced and multifaceted, some details are not addressed in this section.

The primary functions of the bill are as follows. For the overall objective of avoiding AI hazards, this bill sets up regulation for large models on the basis of a positive safety recommendation. To this end, the Frontier Model Division is established. Additionally, a public computing cluster is established: CalCompute.

2.1 Regulation of Large Models

2.1.1 Hazardous Capabilities 22602 (n)

This section quantifies concerns which the bill seeks to address. Hazardous capabilities are defined as a model enabling a mass casualty event, enabling large-scale cyber attack, or autonomously causing great damage through behavior deemed illegal if done by a human.

Crucially, the definition expands to include the potential of posttraining and modification downstream.

This section is well designed, it defines hazardous capability explicitly as relative to what could be accomplished without access to the model. It also establishes only doomsday-esc

scenarios as hazardous. This section is feasible, taking a modest first step in regulating AI. It is effective, by not trying to mitigate every risk, it does not overly hamper innovation.

2.1.2 Covered Models 22602 (f)

This section addresses the scale of model to be considered. The bill applies to large models and uses compute power for training as the main metric to measure the size of the model.

This metric is feasible to measure. However, it is not given that models trained with larger quantities of computing power inherently have greater capacity to cause harm. Conversely, a model trained with less computing power could still be dangerous. This amounts to a possible great design flaw in the bill.

Though, expanding or shifting the scope of the bill could hurt smaller developers and further concentration of power in large companies. So, the current formulation may be the most effective option available.

2.1.3 Positive Safety Determination 22602 (s)

SB 1047 determines the handling of a model on the basis of a positive safety determination. In 22602 (s), a positive safety determination is defined as a determination that a developer can reasonably expect that a model does not and will not possess hazardous capabilities, even with possible posttraining and modification. The notion of a positive safety determination is the bolt upon which the whole bill hinges.

Unfortunately, reliance on words like “reasonably” makes the statement ineffective, it is easy to over or under regulate.

Furthermore, it is not feasible because it breaks the industry it seeks to regulate. There is not currently a method to protect against posttraining overriding safety protocols, so the language of the bill as it stands mandates a halting of all frontier model development.

Thus, the overall design of this section is flawed. The very notion of a positive safety determination is ill defined or unattainable.

2.1.4 Means to Attain Positive Safety Determination

2.1.4.1 Before Training 22603 (a)

By 22603 (a), before training a model, the developer can make a positive safety determination if the model will have lower performance than a non-covered model or than a model that already has a positive safety determination. This protects smaller projects from onerous regulation.

The feasibility of this provision is nice, it relies on accessible metrics and streamlines processes for both State and developer. However, this section is ineffective and poorly designed, it creates a backdoor for unsafe models. If one model is exceptionally safe and gets a positive safety determination, any weaker model has a near automatic positive safety determination.

2.1.4.2 After Training 22603 (c)

If a model was not able to qualify for a positive safety determination before training, 22603 (c) describes a method for attaining one after training: test the model, if it passes, the developer sends a certification to the Frontier Model Division describing the basis for safety determination and test methodology.

This may seem to be ineffective because the developers themselves come up with the tests and could game the system. However, the Frontier Model Division makes the final determination, providing accountability.

This section is well designed because it creates a virtuous cycle: it starts a trend of testing, encourages the development of norms, and will be flexible to new information, research, and findings, allowing those norms to evolve.

Moreover, it is feasible because it delegates the invention of tests to the market and researchers of California, parties who are best suited to develop such tests.

2.1.5 Models Without Positive Safety Determination 22603 (b),(e)&(f)

Where other sections of the bill provide guidance for attaining a positive safety determination, this section of the bill establishes regulation and guidance for the broad category of models that do not yet have a positive safety determination.

22603 (b) (4) requires the developer of a potentially dangerous model to come up with a method for testing and training to make it safe. It serves the purpose of the bill by promoting safety. It is feasible, well designed, and effective because it does not incur any extra cost to the state, but instead demands the cost from the developer, who is creating the hazard in the first place.

Meanwhile, 22603 (b) (9) uses vague language to prohibit training of a model with “unreasonable” risk of hazard. This could have radical chilling effects on model development. If zealously enforced, few (except perhaps large companies) would be able to even initiate training. So, it may be highly effective in terms of mitigating risk, but poorly designed and not feasible for the market.

Finally, 22603 (e)&(f) require periodic safety checkups on models and an annual report to the Frontier Model Division describing the kinds of hazardous capabilities the model has and safety precautions in place. This is feasible, it is low cost for all parties. It is effective and well

designed by requiring an amount of accountability, responsibility, and openness about the dangers that exist with the technology.

2.1.6 Liability 22603 (h)

By 22603 (h), a developer is still liable if the basis for the positive safety determination was unreasonable (ie, failed to account for something that people knew at the time was a risk).

This section is not feasible, poorly designed, and ineffective. It is vague and through it nearly any developer could be held liable for damage done with their model (or any derivative model), regardless of any precautions the developer may have taken.

2.1.7 Computing Clusters 22604

To make regulation of model development possible, 22604 mandates operators of computing clusters gather information on customers using resources sufficient to train a covered model. This information (name, contact information, IP address, etc) will be shared with the Frontier Model Division upon request. This effectively prevents developers from escaping regulation. It is also feasible, many service providers already gather such data about their customers, so it should be easy for computing clusters to enact this requirement.

Arguably this mandates a compromise of privacy and is thus poorly designed, but the standpoint of the bill is that models are potentially very dangerous, so the compromise is justified. Thus, it is well designed.

2.2 Frontier Model Division, Section 4

Section 4 establishes the Frontier Model Division. This division will oversee many provisions set out in this bill, advise other parts of government, and become an authority and guiding body in the field of frontier models.

It is well designed and such a body can be expected to be effective. Unlike a static law, the body can be dynamic, learning and adjusting as frontier models develop.

The feasibility of the division is questionable. It is proposed that they would be funded through fees. It is likely that fees would either be insufficient to fund the Division, or else prohibitively large, thus exacerbating concentration of power among a few successful AI companies.

2.3 CalCompute, Section 5

Section 5 establishes CalCompute, a public cloud computing cluster and starts investigations into what the scope, scale, cost, and appropriate use would be. Thus, the ultimate feasibility and design remains to be seen.

This section is effective because it balances the flaws of the other sections. Regulation in the bill lead to chilling of development and concentration of power in large companies, CalCompute encourages development and democratization.

3. Stakeholder Mapping

The stakeholders that are directly involved in California's SB 1047 represent a very wide array of specialized interests and perspectives.

3.1 Large AI Technology/Research Companies

Large AI Technology/Research companies like Google, Microsoft, OpenAI, Anthropic, Perplexity, etc, are some of the most significant stakeholders. Their rationale for opposition of the bill is that with large-scale regulation and through the establishment of the Frontier Model Division, innovation could be stifled and the velocity with which models can be developed is

slowed down significantly. AI Industry stakeholders would, however, value more clear regulatory frameworks that stabilize the market and maintain a more standardized public trust.

3.2 Startups and Smaller AI Firms

Startups and Smaller AI firms may have a more balanced perspective. Smaller firms might see the bill as a barrier to entry because of imposed compliance costs. This could pose a challenge for smaller AI companies to successfully enter a constantly evolving market. However, the establishment of CalCompute, which operates as a public cloud computing cluster, could give smaller firms reasonable resources to counteract imbalances with larger AI firms. Similar to the large AI companies, there might also be an appreciation for measures that legitimize their development of AI applications and broaden the broader consumer's trust.

3.3 Cloud Infrastructure Providers

Cloud Infrastructure Providers (some of which fall into the category of Large AI technology and research companies) like Amazon Web Services, Microsoft Azure, and Google Cloud would lean towards opposing the bill. Section 22604 outlines customer monitoring and data retention requirements that present a significant liability. This regulatory overhead has the potential for legal exposure for inadvertent hosting/facilitation of unsafe AI models.

3.4 Hardware Manufacturers

Hardware manufacturers (AMD AND NVIDIA) would be generally neutral about this bill/ These manufacturers want to maintain the demand for their hardware products. Their outlook has long anticipated that there would be an increased demand for regulation-compliant computing hardware. These companies are not as directly impacted by the compliance rules laid

forth by the bill, however, they may experience some downstream effects from the broader industry sector's regulation.

3.5 Policy Advocacy Groups

Policy Advocacy Groups (AI Ethics and AI Safety Organizations) are generally in support of the bill. These groups have endorsed a more responsible and rigorous oversight of the development of AI models. Sections 22602 and 22603(c) outline the type of proactive testing that will now be required as safeguards for large-scale AI misuse or damages.

3.6 Civil Liberties Groups

Civil Liberties Groups may take a more cautious stance on this bill. Although they recognize that AI regulation can act in the interest of the public's and individuals' safety, they may exhibit increased concern about potential government overreach.

3.7 General Public

General public opinion will vary widely depending on their stance on AI development. A large segment of the public will appreciate efforts to ensure that AI development remains responsible and safe and does not put their welfare at risk. Others in the public, who are more bullish on the deployment of AI models might cite overregulation as stifling the potential of AI to make breakthroughs in fields like medicine, security, education, etc.

3.8 AI Research Communities and Academia

The broader AI research community and Academia tend to largely support mandates on transparency and other reporting obligations laid out by these regulatory measures. Researchers

and Academia often have the strongest insight into the state of technologies and vulnerabilities that could spur AI misuse, so many academics would lean towards support of the bill. That said, some academics who favor innovation and open research collaboration could view this bill as adversely affecting the state of the art and as an act of overregulation.

4. Potential Weaknesses & Recommended Improvements

This section identifies the major weaknesses of the proposal and proposes specific recommendations.

4.1 Covered Model Definition

The bill defines "covered models" as AI systems trained with computing power exceeding 10^{26} operations or having "similar general capability." This threshold risks becoming outdated as technology evolves. Also, the bill's definition focuses on the computational scale of technologies, but a model trained with less computing power could still be dangerous.

Our recommendation is: revise the "covered model" definition. The bill should abandon its existing approach and instead use a multi-factor framework. This framework should include (1) a dynamic threshold system that is constantly updated by the Frontier Model Division in cooperation with relevant stakeholders, as well as (2) particular benchmarking criteria for analyzing risk capacities across standardized testing.

4.2 Positive Safety Determination

The bill defines positive safety determination as the idea that a developer can "reasonably exclude the possibility" that a model has hazardous capabilities. However, after a safety incident occurs, it's easy to argue that the risk should have been anticipated. The fact that it happened

proves that it was reasonable to expect. This vague bar makes it overly easy to litigate against developers.

Our recommendation is: come up with a standard-based positive safety determination. The bill should provide standards for positive safety determination based on adherence to safety protocols and risk mitigation strategies, requiring developers to provide evidence of safeguards rather than focusing solely on proving a negative.

Furthermore, the bill states that a developer may make a positive safety determination if the covered model performs lower on all benchmarks than one already deemed safe. Here is the weakness: say one powerful model goes through lots of work and effort to make sure that their model is safe. Then, by 22603 (a), any model that is not as powerful has a nearly automatic positive safety determination, even if they put zero effort into safety, or worse, deliberately train with data to make the model capable of critical harm.

Our recommendation is: change the safety determination standard to require more than just lower performance on benchmarks, this is not a reliable proxy for safety. Instead, require that developers demonstrate adherence to safety standards.

4.3 Pre-Training Safety Requirements

For models without a positive safety determination, the bill mandates third-party expert involvement in capability testing and requires that developers implement "the capability to promptly enact a full shutdown"(22603 (b)). However, mandating third-party expert involvement is hard to implement, given that the bill doesn't fully specify how these experts will be identified, engaged, and funded. Also, requiring an immediate shutdown protocol fails to recognize that AI systems are increasingly integrated into critical infrastructure and services, a "prompt" shutdown can have negative consequences.

Our recommendation is: For third-party groups, the bill needs to establish a clear certification process. Regarding shutdown protocols, the bill should adopt a more nuanced containment and mitigation strategy; for example, it should differentiate between settings where full shutdown is possible, like in an isolated research lab, versus settings where full shutdown is less feasible.

4.4 Frontier Model Division Resource Concerns

In Section 4, the bill establishes the Frontier Model Division. This section has two main weaknesses. First, it assigns a great deal of responsibilities to the Frontier Model Division without specifying funding sources and long-term viability. Second, the section is vague, leaving room for abuse on both the enforcement side and the adherence side.

Our recommendation is: first and foremost, ensure consistent funding for the Frontier Model Division — specify a plan for funding that includes budget allocations rather than relying on variable fees. Also, the bill should ensure that it has clear procedural requirements for the Division’s decision-making responsibilities.

5. Reflections

5.1 Matthew Torre:

I was primarily focused on crafting a stakeholder analysis that concisely but thoroughly maps out the motivations and perspectives of large players in the AI space. I read through SB 1047: “Safe and Secure Innovation for Frontier Artificial Intelligence Models Act” and mapped the important policy points to the sentiments of large scale AI companies, startups and smaller AI firms, cloud infrastructure providers, hardware manufacturers, civil liberties groups, policy

advocacy groups, research and academia, and the general public. I tried to take a balanced view of where each stakeholder would position themselves in the spectrum of responsible AI development and potential for overregulation.

5.2 Remington Graham:

I focused on writing the Analysis of Main Components section (and ended up having some analysis that went into the Potential Weaknesses). SB 1047 is a multifaceted bill. It was a great challenge to describe all the dramatic actions it lays out, while keeping within the word count limit laid out in the assignment - especially as reading through the bill, I developed an appreciation for all the little subtleties of each component.

I tried to get an initial overview of the bill written up as quickly as I could to give my groupmates an accessible text to aid in understanding the bill - which by necessity of precision is written with difficult to read language.

It really was a joy to dive into the policy and try to understand the nuance and interplay of each component. Reflecting on the strengths, weaknesses, opportunities, and threats of each aspect gave me a reverence for the difficulty of crafting effective policy.

5.3 Kyran Romero:

Focusing on the theory of impact section of the memorandum, I was able to provide a high-level motivation and goals underlying the development of this bill. Acknowledging intent and motivations behind legislation can help new adopters understand the reasons behind a policy's existence without having to take on the in-depth analysis of the details of the policy beforehand. I was excited to spearhead this preface of our analysis as I naturally find myself thinking behind the ulterior motives behind many of the systems implemented in our societies.

Being that I am studying human-centered AI, I am always very excited to have opportunities to practice the ability to explain why we need to partake in certain actions to regulate or mitigate risks of large AI systems, so this component was very fitting to my skillset!

5.4 Shuvi Jha:

I focused on writing the Potential Weaknesses and Recommended Improvements section. I really enjoyed the opportunity to read SB 1047 and make space to ask myself how I felt about each specific section of the proposal. It felt like a very active and engaging task, one I don't get the chance to do very often in my other coursework.