July 23, 2024

Honorable Buffy Wicks
Chair, California Assembly Appropriations Committee
1021 O Street, Suite 8220
Sacramento, CA 95814

RE: SB 1047 (Wiener) - SUPPORT IF AMENDED

Dear Chair Wicks:

Anthropic is an AI safety and research company based in San Francisco, California. We build among the most powerful AI systems in the world and deploy them to millions of people. We also carry out frontier research into understanding the safety of our AI systems and how to increase their reliability, interpretability, and steerability, and extensively publish our research to benefit the broader ecosystem. We are also proud to contribute to the United States' competitive edge in AI and believe it is important to ensure its continued innovation and leadership. We employ hundreds of people in California and more in other states, Canada, and the United Kingdom, and care deeply about making AI technology that broadly benefits people and society.

Anthropic's position on SB 1047
Anthropic does not support SB 1047 in its current form. However, we believe the bill's core aims to ensure the safe development of AI technologies are worthy, and that it is possible to achieve these aims while eliminating most of the current bill's substantial drawbacks, as we will propose here. In this Support If Amended letter, we outline our views on the risks from AI, the feasibility of safety measures, including those outlined in the bill, and our concerns regarding its current text. We list a set of substantial changes that, if made, would address our multiple concerns and result in a streamlined bill we could support in the interest of a safer, more trustworthy AI industry. Specifically, this includes narrowing the bill to focus on frontier AI developer safety by (1) shifting from prescriptive pre-harm enforcement to a deterrence model that incentivizes developers to implement robust safety and security protocols, (2) reducing potentially burdensome and counterproductive requirements in the absence of actual harm, and (3) removing duplicative or extraneous aspects.

Preserving the safety core of SB 1047

Anthropic takes very seriously the potential for catastrophic risks in future AI systems – both large-scale misuse of AI systems (for example in cyberattacks or bioweapons production) and the potential for AI systems to act autonomously in destructive ways. Although present-day AI systems are in most cases not yet capable of such harm, the trend of capabilities suggests a strong likelihood that they will be in the near future, as our CEO testified in the Senate last year. We do not believe that the private sector has adequate incentives to manage these risks effectively, and thus we believe regulatory incentives to invest in safety and security are needed.

At a high level, SB 1047 attempts to address catastrophic risks by mandating that frontier model developers implement what it calls Safety and Security Protocols ("SSPs"). These appear broadly similar in spirit to frameworks voluntarily adopted by a number of companies, including Anthropic's Responsible Scaling Policy (RSP), OpenAI's Preparedness Framework, Google DeepMind's Frontier Safety Framework, and Magic's AGI Readiness Policy, and which 16 global companies, including Anthropic, committed to publish at the May 2024 AI Seoul Summit.

These frameworks assess each successive generation of AI models for the capability to cause catastrophic harm, and call for the adoption of heightened safety and security measures when risk thresholds are met. Anthropic is a pioneer and strong supporter of such policies. We have now applied our RSP to two major product launches and found it to be both an effective means of managing the risks of developing increasingly capable AI systems, and a reasonable burden for companies with the resources to train a "covered model" as defined in SB 1047 (i.e. a model that costs more than $100M to train).

Based on these experiences, Anthropic is supportive of the idea that frontier AI companies should have some SSP-like plan in place (as noted the majority already do), that the public deserves transparency and honesty about how those plans work, and that frontier AI companies should have an incentive to make those plans actually effective at preventing catastrophic risks. We refer to this as the bill's "safety core".

Summary of Anthropic's proposed framework

We believe the broad opposition to the bill to date is attributable in large part to provisions that are neither essential to this safety core nor even related to it in some instances. We propose to distill the bill to its safety core in a way that could broaden support for it while constituting a substantial leap forward in AI safety. Given the parallels between the safety core and the flexible safety frameworks leading AI labs have independently adopted, we are optimistic that with our proposed amendments the bill will not create an undue burden on AI model developers or impede US competitiveness.

Broadly, we see three types of provisions in the current SB 1047 that are not part of the safety core (and which could be burdensome or antagonize stakeholders) and that we therefore believe should be removed or modified:

- **Broad pre-harm enforcement.** The current bill requires AI companies to design and implement SSPs that meet certain standards – for example they must include testing sufficient to provide a "reasonable assurance" that the AI system will not cause a catastrophe, and must "consider" yet-to-be-written guidance from state agencies. To enforce these standards, the state can sue AI companies for large penalties, even if no actual harm has occurred. While this approach might make sense in a more mature industry where best practices are known, AI safety is a nascent field where best practices are the subject of original scientific research. For example, despite a substantial effort from leaders in our company, including our CEO, to draft and refine Anthropic's RSP over a number of months, applying it to our first product launch uncovered many ambiguities. Our RSP was also the first such policy in the industry, and it is less than a year old. What is needed in such a new environment is iteration and experimentation, not prescriptive enforcement. There is a substantial risk that the bill and state agencies will simply be wrong about what is actually effective in preventing catastrophic risk, leading to ineffective and/or burdensome compliance requirements.

- **New state agencies and powers.** Implementing pre-harm enforcement in turn necessitates the creation of a new state agency (the Frontier Model Division) and new powers (accorded to the Attorney General and Labor Commissioner) to define and enforce compliance standards. These agencies have a mandate that is broad yet vague: to define standards in a technically complex, fast-moving field without established best practices. They also lack firsthand experience developing frontier models. Many stakeholders reasonably worry that this could create an unpredictable situation, where the FMD has very wide latitude and, depending on its opinions or political agenda, might end up harming not just frontier model developers but the startup ecosystem or independent developers, or impeding innovation in general.

- **Provisions without major benefits for catastrophic risk reduction.** These include provisions on pricing, labor law provisions unnecessary to achieve the bill's core safety aims, and customer data collection requirements for cloud service providers that duplicate existing federal requirements.

We believe the first two issues can be addressed[1] by focusing on *deterrence* rather than pre-harm enforcement: instead of deciding what measures companies should take to prevent catastrophes (which are still hypothetical and where the ecosystem is still iterating to determine best practices), focus the bill on holding companies responsible for causing actual catastrophes.

This deterrence approach has a number of advantages. It directly incentivizes companies to prevent catastrophes, and leverages developers' specialized knowledge and empirical insights into their rapidly evolving technologies, allowing for more agile and informed risk management strategies than static regulations might provide. It eliminates the need for new state agencies and rulemaking, resolving much of the ambiguity and complexity of the current bill. And finally, it should appeal to honest skeptics of catastrophic risk, who can choose not to mitigate against risks they don't believe in (though they do so at their own peril).

Our proposed framework is the following:

1. Companies developing models that cost more than $100M to train should be required to establish and publish an SSP, but the state should not prescribe or enforce what is in it – companies should be free to iterate and learn. Best practices will emerge from this process over time.
2. Pre-harm enforcement is limited to honesty and transparency: making sure companies publish their SSP and do not misrepresent their adherence to it.
3. However, if an actual catastrophic incident occurs, and a company's SSP falls short of best practices or relevant standards, *in a way that materially contributed to the catastrophe*, then the developer should also share liability, even if the catastrophe was partly precipitated by a downstream actor.

The third point is the core of the proposal and gets at one of the key ways AI is different from other technologies. If a terrorist used a laptop to plan and execute a mass attack, courts would (rightly) not hold the laptop manufacturer responsible, as the manufacturer realistically has no ability to prevent this misuse. But it's one of our core contentions that AI systems are intelligent and can in principle be trained to avoid engaging in destructive behavior or being appropriated for destructive behavior. This implies that those who train a model owe a duty to investigate potentially catastrophic downstream uses, and to take reasonable measures to prevent such uses. The bill can thus be thought of as mostly a way of strengthening and clarifying existing tort law to focus on this unusual aspect of very powerful AI systems. SSP's can then be seen as public artifacts that document companies'

[1] We address the third issue simply by removing the relevant provisions.

beliefs about what constitutes a reasonable standard of care, and that evolve as it becomes clear where threats truly lie.

Detailed list of changes to the bill

To implement the deterrence vision described above and to pare the bill down to its safety core, we propose the following specific changes:

- **Greatly narrow the scope of pre-harm enforcement** to focus solely on (a) failure to develop, publish, or implement an SSP (the content of which is up to the company); (b) companies making materially false statements about an SSP; (c) imminent, catastrophic risks to public safety.
- **Introduce a clause stating that if a catastrophic event does occur (which continues to be defined as mass casualties or more than $500M in damage), the quality of the company's SSP should be a factor in determining whether the developer exercised "reasonable care."** This implements the notion of deterrence: companies have wide latitude in developing an SSP, but if a catastrophe happens in a way that is connected to a defect in a company's SSP, then that company is more likely to be liable for it.
- **Eliminate the Frontier Model Division (Section 11547.6)**. With pre-harm enforcement sharply limited and no longer prescriptive about standards, the FMD is no longer needed. This greatly reduces the risk surface for ambiguity in how the bill is interpreted, and makes its effects more objective and predictable. In lieu of having an FMD, assign authority to the Government Operations Agency to raise the threshold (initially 10^26 FLOPS and >$100M) for covered models through a notice and comment process to further narrow the scope of covered models as we learn more about risk and safety characteristics of large models over time.
- **Eliminate Section 22605 (uniform pricing for compute and AI models),** which is unrelated to the primary goal of preventing catastrophic risks. It may have unintended consequences for market dynamics in the AI and cloud computing sectors.
- **Eliminate Section 22604 (know-your-customer for large cloud compute purchases),** which duplicates existing federal requirements and is outside the scope of developer safety.
- **Narrow Section 22607 to focus on whistleblowing by employees** that relates to false statements or noncompliance with the company's SSP. Whistleblowing protections make sense and are common in federal and state law, but the language as drafted is too broad and could lead to spurious "whistleblowing" that leaks IP or disrupts companies for reasons unrelated or very tenuously related to catastrophic risk. False statements about an SSP are the area where proactive enforcement remains in our proposal, so it is logical that whistleblower protections focus on this

area in order to aid with enforcement. The proposed changes are in line with, and are not intended to limit, existing whistleblower protections under California's Labor Code.

We also call for a number of minor changes, such as:

- Lowering the expectations for completely precise and independently reproducible testing procedures. Our experience is that policies like SSPs are wet clay and companies are still learning and iterating rapidly on them - if we are overly prescriptive now, we risk "locking the industry in" to poor practices for the long-term. As frontier model training runs may last several months, it is also impractical to state comprehensively and reproducibly the details of all predeployment tests that will be run *before* initiating a months-long training run.
- Removing a potential catch-22 where existing bill text could be interpreted as preventing external testing of a model before a model was tested.
- Removing mentions of criminal penalties or legal terms like "perjury" which are not essential to achieving the primary objectives of the legislation.
- Modifying the "critical harms" definition to clarify that military or intelligence operations in line with the national security objectives of the United States are excluded, and also to remove a vague catch-all critical harm provision. This prevents a company from being liable for authorized government use of force. There is room for debate about the use of AI for military and intelligence objectives. However, we believe the federal level, where responsibility lies for foreign and defense policy, rather than state governments, is the more appropriate forum for such a debate.
- Requiring developers of covered models (>$100M) to publish a public version of an SSP, redacted as appropriate, and retain a copy for five years, in place of filing SSPs (and various other documents) with the FMD (which we have proposed eliminating, as noted above).
- Removing all whistleblower requirements that refer to "any contractor or subcontractor" of the developer of a covered model. This would seem to include anything from data labelers to food vendors. We do not think this bill should introduce new requirements to such a wide swath of businesses, covering thousands to potentially hundreds of thousands of contractors and the contract company employees at large developers. The bill should focus on the direct employees of model developers. Existing whistleblower protections in the Labor Code only extend to employees.
- Setting a more objective and predictable threshold for fine-tuning a covered model to become a new model: $10M or 10% of the cost of the original model, whichever is greater. This level of fine-tuning on top of another developer's model is rare, so this provision limits the compliance requirements of the bill to larger, better resourced

companies. Nevertheless, the language on liability recognizes that fault may be shared across the players in the ecosystem.

- Replacing deletion of model weights as a remedy with the concept of "Full Securing" - deleting all copies of the weights except for a minimal number stored to a high security standard consistent with best practices. Deletion is an extreme remedy, as it could destroy an asset that took enormous time and resources to create. Full securing is a more balanced approach, allowing for enhanced safety measures without the irreversible loss of valuable AI resources.
- Increasing the time for reporting safety incidents from 72 hours to 15 days. Based on our experience deploying AI models, a deadline as short as 72 hours is more likely to distract people trying to respond to a potential incident, than it is to be helpful.

Support if Amended

With these changes, the resulting bill would tightly focus on reducing catastrophic risks through the deterrence approach outlined above. It would respect the evolving nature of risk reduction practices while minimizing rigid, ambiguous, or burdensome rules, preserving a dynamic environment for innovation and U.S. competitiveness. We hope this would lead to a wider range of stakeholders supporting this substantially revised SB 1047.

We are committed to supporting the bill if all of our proposed amendments are made. If a subset of the amendments are agreed to by the final committee, we may or may not support the bill. We will not support the bill as it currently stands.

We are optimistic that if the proposed amendments are adopted it will catalyze an era of innovation and experimentation in risk reduction practices, where companies have skin in the game and are thus incentivized to adopt the practices most likely to actually prevent catastrophic risks.

Sincerely,

Hank Dempsey
State and Local Policy Lead
Anthropic, PBC

CC: The Honorable Scott Wiener, Chair of the California Senate Budget Committee