ANTHROP\C

548 Market St., PMB 90375
San Francisco, CA 94104-5401

412-837-9797
anthropic.com

August 21, 2024

Governor Gavin Newsom
1021 O Street, Suite 9000
Sacramento, CA 95814

RE: SB 1047 (Wiener)

Dear Governor Newsom:

As you may be aware, several weeks ago Anthropic submitted a Support if Amended letter regarding SB 1047, in which we suggested a series of amendments to the bill. Last week the bill emerged from the Assembly Appropriations Committee and appears to us to be halfway between our suggested version and the original bill: many of our amendments were adopted while many others were not.

In our assessment the new SB 1047 is substantially improved, to the point where we believe its benefits likely outweigh its costs. However, we are not certain of this, and there are still some aspects of the bill which seem concerning or ambiguous to us.

In the hopes of helping to inform your decision, we lay out the pros and cons of SB 1047 as we see them, and more broadly we discuss what we see as some key principles for crafting effective and efficient regulation for frontier AI systems based on our experience developing these systems over the past decade.

Pros and Cons of SB 1047
We want to be clear, as we were in our original SIA letter, that SB 1047 addresses real and serious concerns with catastrophic risk in AI systems. AI systems are advancing in capabilities extremely quickly, which offers both great promise for California's economy and substantial risk. Our work with biodefense experts, cyber experts, and others shows a trend towards the potential for serious misuses in the coming years – perhaps in as little as 1-3 years.

We believe SB 1047, particularly after recent amendments, likely presents a feasible compliance burden for companies like ours, in light of the importance of averting catastrophic misuse. The bill's "thresholds" for covered models have been extensively debated, but it's clear the focus is on models requiring substantial development resources. This would affect larger entities like Anthropic and other large frontier labs, while largely exempting smaller players. Our initial concerns about the bill potentially hindering innovation due to the rapidly evolving nature of the field have been greatly reduced in the amended version.

We see the primary benefits of the bill as follows:

- **Developing SSPs and being honest with the public about them**. The bill mandates the adoption of safety and security protocols (SSPs), flexible policies for managing catastrophic risk that are similar to frameworks adopted by several of the most advanced developers of AI systems, including Anthropic, Google, and OpenAI. However, some companies have still not adopted these policies, and others have been vague about them. Furthermore, nothing prevents companies from making misleading statements about their SSPs or about the results of the tests they have conducted as part of their SSPs. It is a major improvement, with very little downside, that SB 1047 requires companies to adopt *some* SSP (whose details are up to them) and to be honest with the public about their SSP-related practices and findings.
- **Deterrence of downstream harms through clarifying the standard of care.** AI systems are more adaptable, intelligent, and programmable than most other general-purpose technologies, and we believe SSP-like measures by foundational AI companies (such as Anthropic) can greatly mitigate the risk of downstream misuse. SB 1047 clarifies existing tort law to ensure that these foundational AI companies may share responsibility for foreseeable downstream risks, and connects their liability to the quality of their SSP. This creates a set of incentives that should encourage companies to develop SSPs that are effective in preventing catastrophic risks. As a company developing foundational models that also invests heavily in safety, Anthropic thinks it is important to systematize and incentivize this attitude across the industry.
- **Pushing forward the science of AI risk reduction.** A number of parties (including Anthropic) have pointed out that AI safety is a nascent field where best practices are the subject of original scientific research. On one hand, this is a reason not to legislate too prescriptively, too early. On the other hand, it highlights the benefits of pushing AI companies to invest in developing this science rather than exclusively investing in making their AI systems more capable and powerful. By requiring

Safety and Security Protocols and clarifying that they will be considered when assessing liability for harms, the bill creates real incentives for companies to take seriously the question of what foreseeable risks their models might be associated with, and how they can build roadmaps to having appropriate risk mitigations by the time they are imposing potential risks on society.

Taken together, these aspects of SB 1047 have the potential to meaningfully improve our ability to prevent serious risks from AI systems.

That said, we still have concerns about some aspects of SB 1047, and it is worth laying these out:

- **Some concerning aspects of pre-harm enforcement are preserved in auditing and GovOps.** One of our central concerns in our "Support if Amended" letter was the Frontier Model Division's (FMD) prescriptive guidance, reinforced by pre-harm enforcement. We believe this approach was too rigid, given the nascent state of AI technology. In the amended SB 1047, the FMD is eliminated and pre-harm enforcement is substantially narrowed, but some of the FMD's powers have been moved to GovOps, and GovOps can now issue binding requirements for private auditors. The interplay of these entities is complex: GovOps issues non-binding guidance on SSPs, but also shapes the conditions for audits, which are mandatory to perform (though not mandatory to adopt the recommendations of). In addition, auditors are tasked with measuring compliance with *all* requirements of the bill, which includes language about "reasonable care" to avoid harm. It is our best understanding that this interplay will *not* end up causing unnecessary pre-harm enforcement, but the language has enough ambiguity to raise concerns. If implemented well, this could lead to well-defined standards for auditors and a well-functioning audit ecosystem, but if implemented poorly this could cause the audits to not focus on the core safety aspects of the bill.
- **The bill's treatment of injunctive relief.** Another place pre-harm enforcement still exists is that the Attorney General retains broad authority to enforce the entire bill via injunctive relief, including before any harm has occurred. This is substantially narrower than previous pre-harm enforcement, but is still a vector for overreach.
- **Miscellaneous other issues.** A number of other issues raised in our "Support if Amended" letter, such as know-your-customer requirements on cloud providers,

overly short notice periods for incident reporting, and overly expansive whistleblower protections that are subject to abuse, were not addressed.

The burdens created by these provisions are likely to be manageable, *if* the executive branch takes a judicious approach to implementation. If SB 1047 were signed into law, we would urge the government to avoid overreach in these areas in particular, to maintain a laser focus on catastrophic risks, and to resist the temptation to commandeer SB 1047's provisions to accomplish unrelated goals.

Thoughts on Regulating Frontier AI Systems
Regardless of whether or not SB 1047 is adopted, California will be grappling with how to regulate AI technology for years to come. Below we share our general perspective on AI regulation, which we hope may be useful in considering both SB 1047 and future regulatory efforts that might occur instead or in addition to it.

First some high-level principles:

- **The key dilemma of AI regulation is driven by speed of progress.** AI technology continues to advance extremely rapidly. On one hand, this means that regulation is urgently needed on some key issues: we believe that these technologies will present serious risks to the public in the near future. On the other hand, precisely because the field is advancing so quickly, strategies for mitigating risk are in a state of rapid evolution, often resembling scientific research problems more than they resemble established best practices. We believe that this genuinely difficult dilemma is one important driver of the divergence in views among different AI experts on SB 1047 and in general.
- **One resolution to this dilemma is very adaptable regulation.** In grappling with the dilemma above, we've come to the view that the best solution is to have a regulatory framework that is very adaptable to rapid change in the field. There are several ways to accomplish this, including via third party auditors, frameworks that shape incentives without prescribing behavior, or procedural requirements that require a safety *process* without prescribing what is in it. Down the road (perhaps in as little as 2-3 years), when best practices are better established, a prescriptive framework could make more sense – prescriptive frameworks often work in mature industries such as aerospace or automobiles.

- **Catastrophic risks are important to address.** AI obviously raises a wide range of issues, but in our assessment catastrophic risks are the most serious and the least likely to be addressed well by the market on its own. As noted earlier in this letter, we believe AI systems are going to develop powerful capabilities in domains like cyber and bio which could be misused – potentially in as little as 1-3 years. In theory, these issues relate to national security and might be best handled at the federal level, but in practice we are concerned that Congressional action simply will not occur in the necessary window of time. It is also possible for California to implement its statutes and regulations in a way that benefits from federal expertise in national security matters: for example the NIST AI Safety Institute will likely develop non-binding guidance on national security risks based on its collaboration with AI companies including Anthropic, which California can then utilize in its own regulations.

In terms of specific properties of an AI frontier model regulatory framework, we see three key elements as essential:

1. **Transparent safety and security practices.** At present, many AI companies evidently consider it necessary to have detailed safety and security plans for managing AI catastrophic risk, but the public and lawmakers have no way to verify adherence to these plans or the outcome of any tests run as part of them. Transparency in this area would create public accountability, accelerate industry learning, and promote a "race to the top," with very few downsides. Many different mechanisms might be used to create transparency; what matters is the end result.
2. **Incentives to make safety and security plans effective in preventing catastrophes.** Point 1 alone could lead to a situation where companies can declare very weak safety and security practices while facing only the very soft incentive of public disapproval. It seems important to supplement this with harder incentives. As stated in our initial SIA letter, we believe AI companies are currently better positioned than most other actors to figure out which practices are most effective at preventing risk, so incentivizing the right *outcome* seems more promising than prescribing rules. There are several potential mechanisms for doing this, including strengthening liability for catastrophes, creating a system of private regulators who are incentivized to prevent catastrophes, or through regulation of insurance. Also, as the industry becomes more mature, prescriptive rules may gradually become more appropriate.

3. **Minimize collateral damage.** AI catastrophic risk is an emerging field, where there is great room for disagreement and expert and industry opinions are in flux. One of the worst things that could happen to this field is to create an association between regulation to prevent these risks, and burdensome or illogical rules. It is important, in general but especially in this case, for regulation to be as “clean” as possible, incurring only the burdens absolutely necessary to prevent risk while carefully avoiding collateral damage. We believe that SB 1047 has accumulated as much opposition as it has in part because the bill’s proponents underrated this issue until very late in the process.

We believe it is critical to have *some* framework for managing frontier AI systems that roughly meets these three requirements. As AI systems become more powerful, it's crucial for us to ensure we have appropriate regulations in place to ensure their safety.

Sincerely,

Dario Amodei
Chief Executive Officer
Anthropic, PBC