Part 1: Organization Trust Analysis

I. Introduction

When Anthropic's seven co-founders left OpenAI in 2021, they positioned their new company as the responsible alternative. A Public Benefit Corporation that would prioritize AI safety over profits.[1] The founding narrative was compelling: frustrated by OpenAI's Microsoft partnership and scaling-first approach, these elite researchers would build AI "for the long-term benefit of humanity." [2] Within three years, Anthropic secured over $11 billion from Amazon and Google[3] (more on Valuation in Exhibit 03), released Claude models that competed at the capability frontier, and established the industry's most sophisticated trust architecture including Constitutional AI[4], a Responsible Scaling Policy[5], and an independent Long-Term Benefit Trust with board control[6].

However, between February and September 2024, California's SB-1047 AI safety bill[7] exposed a critical gap between Anthropic's safety rhetoric and its lobbying practice. While publicly expressing cautious support, the company privately worked to weaken whistleblower protections, eliminate pre-harm enforcement, and introduce procedural obstacles. Problems and uncertainty are opportunities that have saved Anthropic's reputation amongst the competitors. This episode demonstrates how competitive pressure can erode even the most carefully

[1] *Wikipedia*, "Anthropic," December 5, 2025, https://en.wikipedia.org/w/index.php?title=Anthropic&oldid=1325800170.
[2] "Report: Anthropic Business Breakdown & Founding Story | Contrary Research," accessed December 15, 2025, https://research.contrary.com/company/anthropic.
[3] Hayden Field, "Amazon to Invest Another $4 Billion in Anthropic, OpenAI's Biggest Rival," CNBC, November 22, 2024, https://www.cnbc.com/2024/11/22/amazon-to-invest-another-4-billion-in-anthropic-openais-biggest-rival.html.
[4] "Constitutional AI: Harmlessness from AI Feedback," accessed December 15, 2025, https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback.
[5] "Anthropic's Responsible Scaling Policy," accessed December 15, 2025, https://www.anthropic.com/news/anthropics-responsible-scaling-policy.
[6] "The Long-Term Benefit Trust," accessed December 15, 2025, https://www.anthropic.com/news/the-long-term-benefit-trust.
[7] "SB 1047- ENROLLED," accessed December 15, 2025, https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB1047.

constructed organizational trust, raising fundamental questions about whether a safety-first mission can survive in frontier AI.

II. Anthropic's Trust-Building Framework (400 words)

Competence: *Technical Excellence and Elite Team*

Anthropic's technical credibility stems from pioneering research and exceptional talent. The Constitutional AI paper (December 2022) revolutionized alignment methodology by training systems through explicit principles rather than extensive human feedback[8]. The interpretability team, led by Chris Olah, achieved the first detailed look inside a production LLM with their "Scaling Monosemanticity" work (May 2024), extracting millions of interpretable features including those for deception and power-seeking.[9]

The team's credentials are exceptional: Dario Amodei led GPT-2/GPT-3 development, co-invented RLHF, and holds 142,371+ citations. Chris Olah essentially created the neural network interpretability field and was named to TIME's 100 Most Influential People in AI (2024)[10]. All seven co-founders remain at the company. Product performance validates this expertise via Claude 3.5 Sonnet achieved 92% on HumanEval (coding benchmark)[11], outperforming GPT-4o,[6] while Claude 4 demonstrated 24-hour autonomous task completion. All major safety innovations are peer-reviewed and publicly available, contrasting with competitors' opacity.

Motives: *Mission-Driven Structure*

Seven co-founders left OpenAI in late 2020-early 2021 after concluding that scaling required

[8] “Constitutional AI.”
[9] “Chris Olah on What the Hell Is Going on inside Neural Networks,” *80,000 Hours*, n.d., accessed December 15, 2025, https://80000hours.org/podcast/episodes/chris-olah-interpretability-research/.
[10] “Chris Olah: The 100 Most Influential People in AI 2024 | TIME,” accessed December 15, 2025, https://time.com/7012873/chris-olah/.
[11] “Claude 3.5 Sonnet Review (Performance & Benchmarks),” accessed December 15, 2025, https://textcortex.com/post/claude-3-5-sonnet.

"something in addition...which is alignment or safety." The departure followed OpenAI's Microsoft partnership, though Amodei attributes it to vision differences rather than conflict.[12]

The corporate structure institutionalizes safety through Delaware Public Benefit Corporation status which legally requires directors to balance shareholder interests with "responsible development and maintenance of advanced AI for the long-term benefit of humanity." The Long-Term Benefit Trust (LTBT, September 2023) provides five financially disinterested trustees with increasing board control as of currently two of five directors, eventually three of five.[13]

Constitutional AI makes values explicit and inspectable, drawing from the UN Declaration of Human Rights. The Responsible Scaling Policy (September 2023) commits to implementing "safety and security measures that will keep risks below acceptable levels" before deploying catastrophically capable models.Amodei's "Machines of Loving Grace" essay (October 2024) frames safety as enabling commercial success, arguing both AI's upside and risks are underestimated.[14]

Means: *Governance Architecture and Vulnerabilities*

The LTBT features trustees like Neil Buddy Shah (Clinton Health Access Initiative CEO) with national security and policy backgrounds who hold no equity stake yet critics note stockholders can rewrite LTBT rules, and the unpublished Trust Agreement leaves real authority unclear[15].

The RSP establishes graduated AI Safety Levels (ASL-2 basic current baseline, ASL-3 enhanced and activated May 2025 for Claude Opus4, ASL-4 undefined till today). Transparency includes bug bounties up to $35,000, model cards and third-party evaluations[16].

---

[12] Gennaro Cuofano, “Who Is Dario Amodei?,” FourWeekMBA, July 22, 2024, https://fourweekmba.com/who-is-dario-amodei/.
[13] “The Long-Term Benefit Trust”; *Wikipedia*, “Anthropic.”
[14] “Dario Amodei — Machines of Loving Grace,” accessed December 15, 2025, https://www.darioamodei.com/essay/machines-of-loving-grace.
[15] Billy Perrigo/San Francisco, “How Anthropic Designed Itself to Avoid OpenAI’s Mistakes,” TIME, May 30, 2024, https://time.com/6983420/anthropic-structure-openai-incentives/.
[16] “Activating AI Safety Level 3 Protections,” accessed December 15, 2025, https://www.anthropic.com/news/activating-asl3-protections.

Employee protections proved problematic: Anthropic's 2024 whistleblower policy came only after exposing concealed non-disparagement agreements that co-founder Sam McCandlish called "unclear" contradicted by ex-employee Neel Nanda confirming his explicitly prohibited disclosure[17].

Impact: *Track Record and Accountability*

Anthropic has maintained a strong safety record. Claude powers integrations in Slack, Notion, and DuckDuckGo with few reported misuse incidents[18]. The company's safety-first positioning has pushed industry standards. CEO Amodei hopes their "existence in the ecosystem causes other organizations to become more like us."[19]

However, with trust comes responsibility. Anthropic has been constantly scrutised as they position themselves as AI Saftey first Company. For example:
SB-1047: Publicly "supported if amended" while privately lobbying against pre-harm enforcement and whistleblower protections tactics Max Tegmark called "straight out of Big Tech's playbook" resulting in a significantly weakened bill that was ultimately vetoed. *This issue we tackle in depth in the next section.*

RSP modifications: Quietly removed October 2023 commitments to "pause in scaling" and define ASL-4 before training ASL-3 models (October 2024), then weakened insider threat protections just before Claude Opus 4 release (May 2025)[20].

[17] MikhailSamin, *Unless Its Governance Changes, Anthropic Is Untrustworthy*, December 2, 2025, https://forum.effectivealtruism.org/posts/6XbtL93kSFJwX45X2/unless-its-governance-changes-anthropic-is-untrustworthy.
[18] *Wikipedia*, “Anthropic.”
[19] “TIME100 AI 2023: Dario and Daniela Amodei,” Time, September 7, 2023, https://time.com/collection/time100-ai/6309047/daniela-and-dario-amodei/.
[20] “Responsible Scaling Policy Updates,” accessed December 15, 2025, https://www.anthropic.com/rsp-updates?subjects=announcements.

Non-disparagement agreements: Failed to proactively disclose identical practices to OpenAI's exposed agreements (May 2024), then falsely characterized them as "unclear" despite employees confirming explicit prohibition on disclosure[21].

Copyright infringement: Settled for $1.5 billion (largest in U.S. history) after using 7 million pirated books from LibGen and PiLiMi[22].

Authoritarian funding: Leaked memo revealed pursuit of UAE/Qatar investments despite Amodei acknowledging this would "enrich dictators."[23]

III. The SB-1047 Crisis

The California Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (SB-1047)[24] controversy of 2024 represents Anthropic's most documented trust crisis, revealing tensions between the company's safety-first branding and its alleged private lobbying tactics. The incident damaged trust across multiple stakeholder groups and produced extensive primary source documentation.

February 7, 2024: Senator Scott Wiener introduces SB-1047, requiring frontier AI companies to implement safety protocols, testing requirements, and face liability for catastrophic harms.

Early 2024: CEO Dario Amodei publicly states regulation is "too early," arguing industry consensus around responsible scaling policies should come first.[25]

---

[21] MikhailSamin, *Unless Its Governance Changes, Anthropic Is Untrustworthy*.
[22] Chloe Veltman, "Anthropic Settles with Authors in First-of-Its-Kind AI Copyright Infringement Lawsuit," Culture, *NPR*, September 5, 2025, https://www.npr.org/2025/09/05/nx-s1-5529404/anthropic-settlement-authors-copyright-ai.
[23] *Wikipedia*, "Dario Amodei," December 14, 2025, https://en.wikipedia.org/w/index.php?title=Dario_Amodei&oldid=1327437590.
[24] *Wikipedia*, "Safe and Secure Innovation for Frontier Artificial Intelligence Models Act," November 16, 2025, https://en.wikipedia.org/w/index.php?title=Safe_and_Secure_Innovation_for_Frontier_Artificial_Intelligence_Models_Act&oldid=1322494960.
[25] Sharon Goldman, "It's AI's 'Sharks vs. Jets'—Welcome to the Fight over California's AI Safety Bill," Fortune, accessed December 15, 2025, https://fortune.com/2024/07/15/california-ai-bill-sb-1047-fierce-debate-regulation-safety/.

Observers noted that Anthropic was “pushing back on a landmark California bill to regulate AI” a move that seemed at odds with its “good guy” reputation[26]. News outlets highlighted the unusual alliances and divides around SB-1047. Axios broke the story that Anthropic did not support the bill’s initial version, even as it touted AI safety goals[27]. Reddit flooded with expressing the loss of trust for a company that advocates for AI Safety and Governance[28]. A detailed **December 2025** analysis on LessWrong and the EA Forum by researcher Mikhail Samin documented allegations that Anthropic engaged in "acoustic separation", communicating different messages to different stakeholders[29]. According to this analysis[30]:

1. Leadership initially attempted to convince employees that state-level regulation would be "terrible" and should be opposed in favor of federal legislation only.
2. When employees pushed back, the company "appeared to somewhat give in and reduced its opposition" .
3. Anthropic allegedly introduced amendments designed to "touch on the scope of every committee in the legislature, thereby giving each committee another opportunity to kill the bill" .
4. Amazon, Anthropic's $8 billion investor, allegedly influenced the company's lobbying against Know Your Customer provisions that would have affected cloud providers.

IV. Turn: World Leader reputation

Within the broader AI sector, Anthropic’s position set it apart from certain peers. OpenAI leadership had often called for AI regulation in theory, but when faced with SB-1047 they lobbied

[26] Sigal Samuel, “It’s Practically Impossible to Run a Big AI Company Ethically,” Vox, August 2, 2024, https://www.vox.com/future-perfect/364384/its-practically-impossible-to-run-a-big-ai-company-ethically.
[27] Ashley Gold, “Exclusive: Anthropic Weighs in on California’s AI Bill,” Axios, July 25, 2024, https://www.axios.com/2024/07/25/exclusive-anthropic-weighs-in-on-california-ai-bill.
[28] katxwoods, “LA Times endorses bill SB 1047, saying that this isn’t the first time big tech leaders have publicly professed that they welcome regulation on their products, but then lobbied fiercely to block specific proposals.,” Reddit Post, R/Singularity, August 23, 2024, https://www.reddit.com/r/singularity/comments/1ez40dj/la_times_endorses_bill_sb_1047_saying_that_this/.
[29] MikhailSamin, *Unless Its Governance Changes, Anthropic Is Untrustworthy*.
[30] “Unless Its Governance Changes, Anthropic Is Untrustworthy,” accessed December 15, 2025, https://anthropic.ml/.

against it aggressively. In August 2024, OpenAI's strategy chief warned Governor Newsom that "SB 1047 would threaten (California's) growth, slow innovation, and lead… engineers to leave the state," urging that AI rules be left to the federal government instead[31]. OpenAI's public letter did not target specific SB-1047 provisions so much as argue no single state should regulate AI. Other competitors [Exhibit 02] were on a very similar tone. However, in **July 2024** "Support If Amended" letter to lawmakers, Anthropic praised the bill's goal of AI safety but warned that "the current version of SB 1047 has substantial drawbacks that harm its safety aspects and could blunt America's competitive edge in AI development." This updated the reputation of Anthropic and gained back confidence. Axios and others noted Anthropic had offered "cautious support" once the bill was revised, a stance contrasted with OpenAI's open opposition[32].

This evolution allowed Anthropic to be seen as a responsible stakeholder rather than an industry antagonist. Indeed, over 100 employees across top AI companies (including Anthropic) signed an open letter urging SB-1047's passage, explicitly countering their own CEOs' objections[33]. The letter signed by Anthropic's co-founder Chris Olah, Turing Award winner Geoffrey Hinton, Yoshua Bengio, and many others – argued that safeguarding against severe AI risks is "feasible and appropriate" and called SB-1047 "a meaningful step forward." The fact that nearly half of the signatories were Anthropic employees (including several who had left OpenAI) was a visible company culture. It suggested that Anthropic's workforce broadly supported the kind of accountability SB-1047 envisioned, which in turn bolstered Anthropic's credibility as an organization genuinely committed to AI safety. This public endorsement by its team, in defiance of broader tech opposition, reflected well on Anthropic's culture and helped differentiate it from OpenAI and Meta.

---

[31] Maxwell Zeff, "OpenAI's Opposition to California's AI Bill 'makes No Sense,' Says State Senator," *TechCrunch*, August 21, 2024, https://techcrunch.com/2024/08/21/openais-opposition-to-californias-ai-law-makes-no-sense-says-state-senator/.
[32] Ina Fried Gold Ashley, "California's AI Safety Bill Is Dividing Big Tech," Axios, August 28, 2024, https://www.axios.com/2024/08/28/california-ai-regulation-bill-divides-tech-world.
[33] "Dozens of AI Workers Buck Their Employers, Sign Letter in Support of Wiener AI Bill," September 9, 2024, https://sfstandard.com/2024/09/09/ai-workers-support-wiener-bill/.

On **August 21, 2024** Anthropic's policy head, Jack Clark and CEO Dario Amodei shared a carefully worded letter illustrating the frustration and solution to the Bill SB-1047 [Exhibit 04]. As one analysis put it, Anthropic – often described as "the AI lab most associated with AI safety" – initially had reservations about SB-1047, but *"after further revisions, [it] concluded that the final bill's benefits outweighed its total costs."*

Senator Wiener praised Anthropic as "a world leader on both innovation and safety" as the bill advanced to the Assembly Floor with amendments. The revisions incorporated Anthropic's suggestions, including provisions to "accommodate the unique needs of the open source community, which is an important source of innovation. [Exhibit 07]

However, on **September 29, 2024** the Governor Gavin Newsom vetoed [Exhibit 06] the bill and and returns SB-1047 without a signature. Also, *The San Francisco Standard* called Anthropic's late-August endorsement of the amended bill *"the first major crack in the AI industry's near-uniform resistance"* to SB-1047[34].

Later in **September 2025**, SB-53 was introduced as the successor to SB-1047, explicitly framed as a course correction. California lawmakers rewrote the bill to address the political and trust failures exposed by SB-1047, especially concerns about overbreadth, premature enforcement, and chilling innovation[35].

SB-53 marked a reputational inflection point. After the backlash around SB-1047, Anthropic appeared to internalize that how a company engages with regulation affects trust as much as what it argues [Exhibit 05] . By accepting SB-53 quietly and cleanly, Anthropic avoided repeating the trust erosion caused by its earlier posture.

[34] "Dozens of AI Workers Buck Their Employers, Sign Letter in Support of Wiener AI Bill."
[35] "Governor Newsom Signs Senator Wiener's Landmark AI Law To Set Commonsense Guardrails, Boost Innovation," Senator Scott Wiener, September 29, 2025, https://sd11.senate.ca.gov/news/governor-newsom-signs-senator-wieners-landmark-ai-law-set-commonsense-guardrails-boost.

V. Conclusion

Anthropic built the industry's most sophisticated trust architecture which has PBC structure, Constitutional AI, Long-Term Benefit Trust, Responsible Scaling Policy. Yet SB-1047 exposed the gap between safety rhetoric and lobbying practice. While 113 employees including co-founder Chris Olah publicly supported the bill, leadership privately weakened whistleblower protections and enforcement mechanisms under Amazon's $8 billion influence (as in news). The pattern repeated through quiet RSP modifications, concealed non-disparagement agreements, and authoritarian funding pursuits. Anthropic remains better than competitors. It is qualified SB-1047 support contrasted with OpenAI's blanket opposition but the fundamental question persists: ***can any frontier AI company sustain mission-driven governance when survival demands capability racing and regulatory weakening?***

Part 2: Personal Reflections on Leading with Trust

"People will say, it is their job to cristise"

The One Liner

Saying is one thing, doing is another!

Power dynamics and their evolution over time may need some adaptations.

Beyond Individual Ethics

Anthropic's trajectory does not invalidate the competence/motives/means/impact framework but it makes me question that these dimensions will not remain stable over time all at once. The framework allowed to dissect the issues and to maintain trust all four has to be in harmony. The company possessed exceptional competence, explicitly stated safety motives, innovative governance means and strong initial impact. Yet trust is fractured or communicated because the framework has to constantly account for how external pressures or power transform these dimensions dynamically. [Exhibit 01]

Hence, there is a ***temporal dimension of trust.***

What I Actually Need to Learn: Recognizing When Constraint Has Failed

The competence/motives/means/impact framework directs attention toward building better governance at founding. But Anthropic's LTBT was innovative governance. The RSP was an industry-leading policy. The PBC structure was legally novel. Will it be able to prevent mission drift? The framework is to be applied at each step. Is it designed to constrain bad-faith actors but not to resist good-faith rationalization under pressure? Deeper study and knowledge will help me grow and learn in this direction.

“We commit” is not equal to “ We generally try”.

Three Commitments, Not Skills

Rather than listing skills to develop, I commit to three practices that might make constraint failure visible:

- Maintain a decision log with stated justifications
- Establish irreversibility triggers so that I can adhere to vision and mission
- Cultivate dissent as structural feature

The question I cannot answer

The assignment asks what I learned about "leading with trust." But Anthropic raises a question I cannot answer: ***Is trust-based leadership possible in frontier industries where survival requires continuous compromise of founding principles?***

Exhibits:

01: Temporal based Anthropics' Competence, Motives, Means and Impact

Year	Event	Competence (Can they do it?)	Motives (Why they do it?)	Means (How they do it?)	Impact (What changed?)
2021	Founded as Public Benefit Corporation	Founders are ex-OpenAI senior researchers	Explicit safety-first break from OpenAI commercialization	Legal structure embeds public benefit	Immediate credibility with AI safety community
2021	$124M Series A (Moskovitz, Tallinn)	Trusted to build frontier models	Backers aligned with long-term risk reduction	Patient capital, not revenue pressure	Ability to prioritize research over product
2022	Claude v1 trained but not released	Demonstrated frontier-scale LLM capability	Avoid triggering unsafe deployment race	Internal-only testing, no public API	Strong signal of restraint, rare in industry
2022	Constitutional AI paper published	Original alignment research contribution	Align models to human rights, not engagement	Transparent, publish-befor e-product	Raised bar for alignment discourse

2022	$580M Series B (FTX-led)	Confidence in scaling capability	Mixed signal due to FTX association	Accepted capital without governance control	Short-term reputational risk, later neutralized
2023	Claude private beta	Stable, usable assistant	Gradual exposure over hype	Limited rollout via partners only	Trust via controlled experimentation
2023	Claude constitution publicly disclosed	Clear value grounding	Ethical transparency	Publicly auditable principles	Differentiation from opaque competitors
2023	White House AI Safety Commitments	Recognized as credible lab	Public-interest alignment	Voluntary external oversight	Policy legitimacy
2023	Amazon investment ($4B)	Proven production readiness	Sustain independence while scaling	LTBT governance shield	Balanced growth with safeguards
2024	Hiring Leike and Schulman (ex-OpenAI)	Deep alignment expertise	Talent chooses safety culture	Strengthened alignment team	Reinforced internal trust
2024	Mechanistic interpretability research	Rare model introspection capability	Reduce black-box risk	Publish neuron-level analysis	Academic credibility boost

2025	National security access restrictions	Awareness of geopolitical misuse	Prioritize harm prevention over revenue	Sales exclusions by ownership	Trust with governments
2025	LTBT gains board control	Governance competence	Mission enforcement over profit	Trustee-majority board	Durable trust architecture

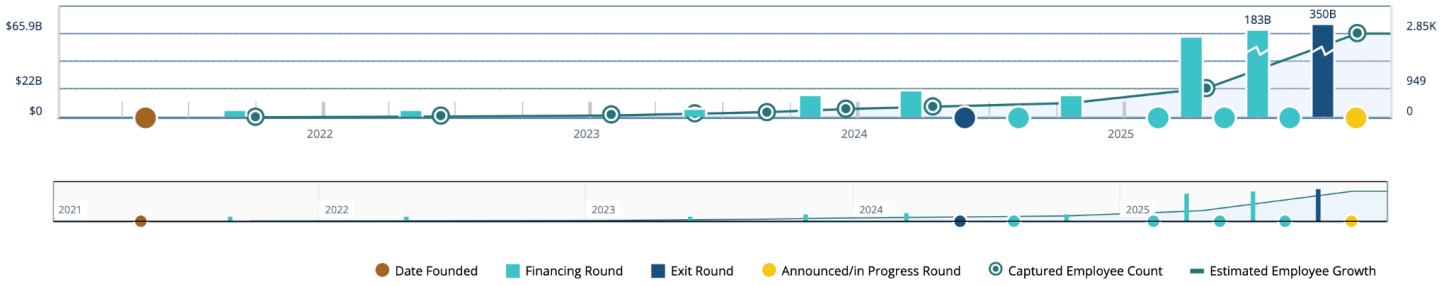
02: Anthropic amongst its competitors(Trust-Oriented Comparison)

Anthropic wins on governance credibility, not on openness or raw capability.

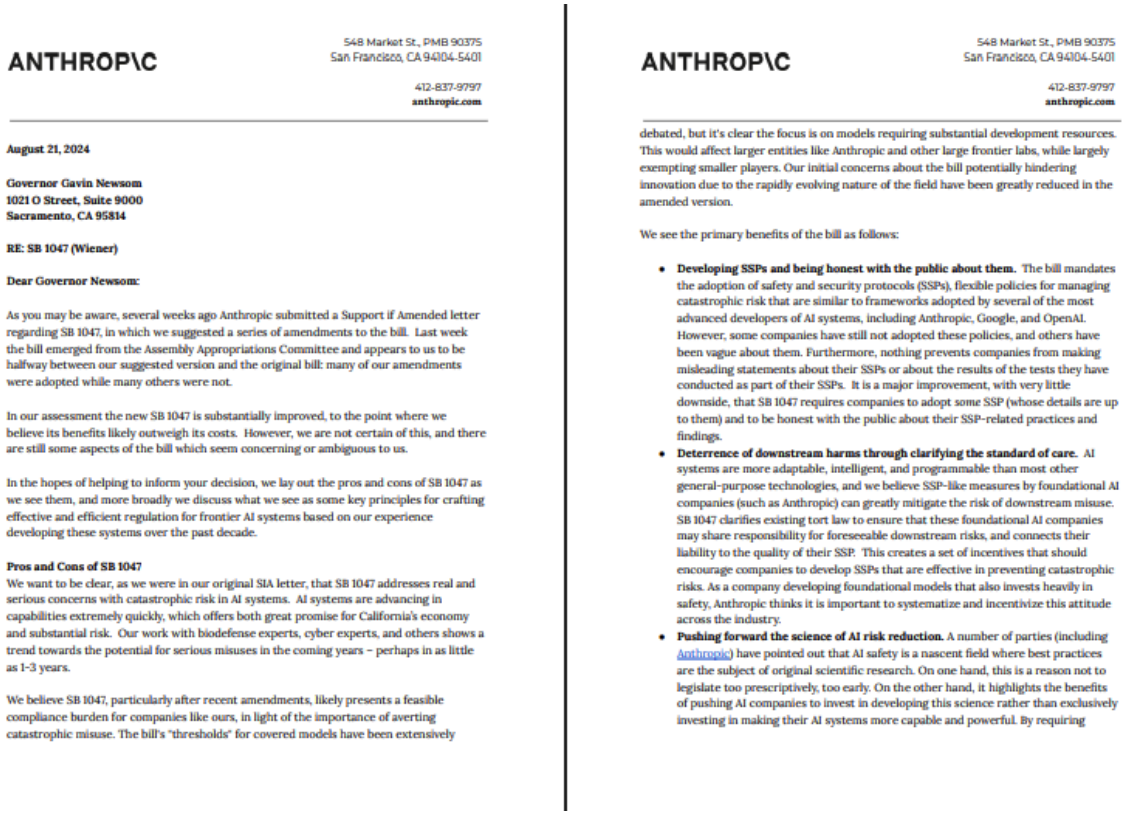
Organization	**Core Strength**	**Trust Posture**	**Transparency Level**	**Governance Model**	**Primary Criticism**
Anthropic	Alignment-first frontier models	Safety-led	High on values, medium on tech	PBC + Long-Term Benefit Trust	Still scaling aggressively
OpenAI	Capability leadership	Product-led	Low post-GPT-4	Capped-profit nonprofit	Governance instability, opacity
Google DeepMind	Research depth	Institutional safety	Medium	Corporate subsidiary	Slower public accountability
Meta AI	Open research	Speed and openness	High research openness	Corporate	Weaker deployment safeguards

xAI	Rapid scaling	Ideology-driven	Low	Founder-controlled	Minimal safety signaling
Mistral AI	Open-weight models	Developer trust	High technical openness	Venture-backed	Less alignment emphasis

03: Anthropic's Round and Valuation Timeline (Source: Pitchbook)



04: Anthropic's Letter to Governor for SB-1047 (Source: Anthropic's website)



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

ANTHROP\C

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

412-837-9797
anthropic.com

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.

ANTHROP\C

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

412-837-9797
anthropic.com

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.

ANTHROP\C

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

412-837-9797
anthropic.com

- **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.

ANTHROP\C

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

412-837-9797
anthropic.com

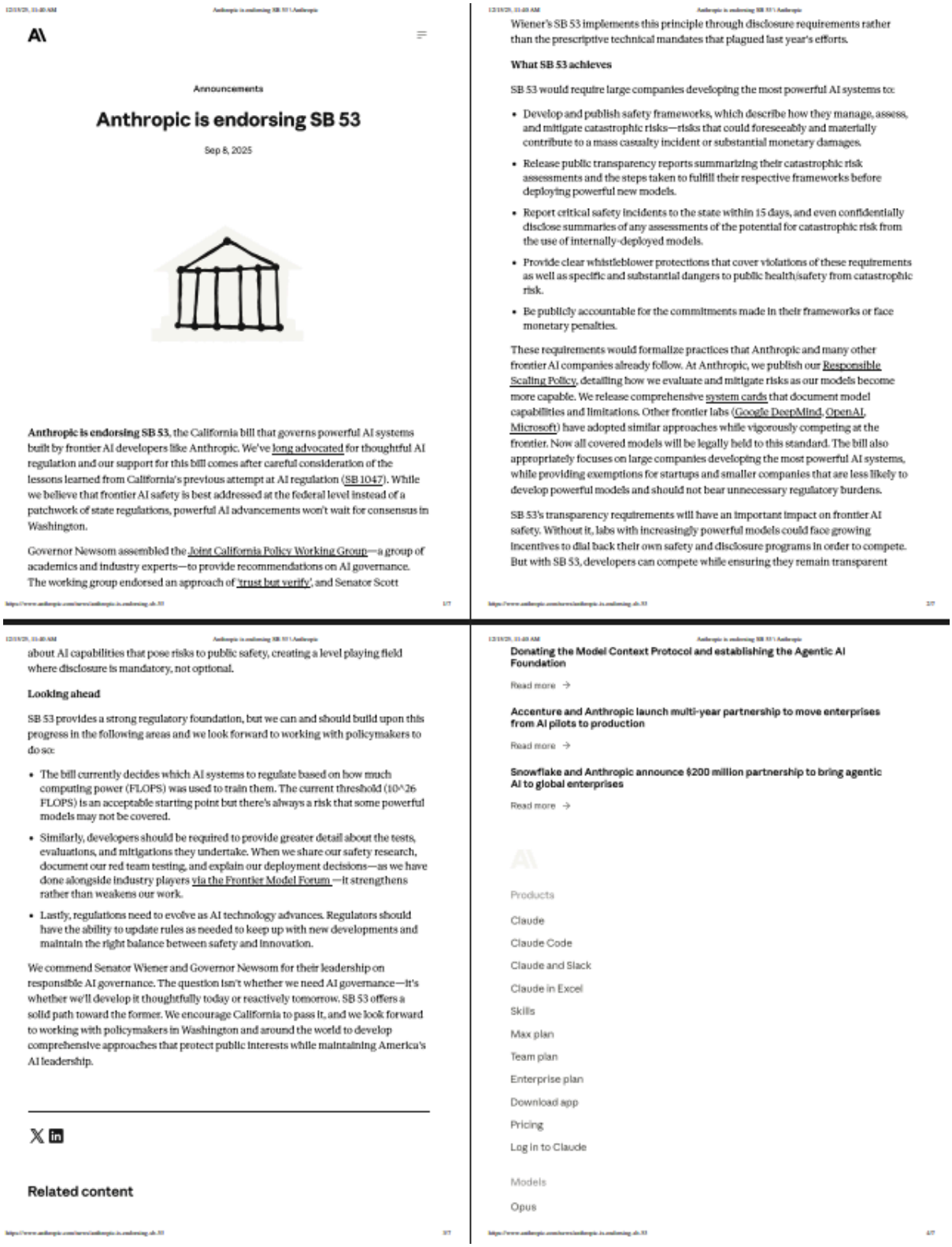
3. **Minimize collateral damage.** AI catastrophic risk is an emerging field, where there is great room for disagreement and expert and industry opinions are in flux. One of the worst things that could happen to this field is to create an association between regulation to prevent these risks, and burdensome 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

05: Anthropic’s endorsement to SB 53 in September 2025 (Source: Anthropic official website)



Announcements

Anthropic is endorsing SB 53

Sep 8, 2025

Anthropic is endorsing SB 53, the California bill that governs powerful AI systems built by frontier AI developers like Anthropic. We've long advocated for thoughtful AI regulation and our support for this bill comes after careful consideration of the lessons learned from California's previous attempt at AI regulation (SB 1047). While we believe that frontier AI safety is best addressed at the federal level instead of a patchwork of state regulations, powerful AI advancements won't wait for consensus in Washington.

Governor Newsom assembled the Joint California Policy Working Group—a group of academics and industry experts—to provide recommendations on AI governance. The working group endorsed an approach of 'trust but verify', and Senator Scott Wiener's SB 53 implements this principle through disclosure requirements rather than the prescriptive technical mandates that plagued last year's efforts.

What SB 53 achieves

SB 53 would require large companies developing the most powerful AI systems to:

- Develop and publish safety frameworks, which describe how they manage, assess, and mitigate catastrophic risks—risks that could foreseeably and materially contribute to a mass casualty incident or substantial monetary damages.
- Release public transparency reports summarizing their catastrophic risk assessments and the steps taken to fulfill their respective frameworks before deploying powerful new models.
- Report critical safety incidents to the state within 15 days, and even confidentially disclose summaries of any assessments of the potential for catastrophic risk from the use of internally-deployed models.
- Provide clear whistleblower protections that cover violations of these requirements as well as specific and substantial dangers to public health/safety from catastrophic risk.
- Be publicly accountable for the commitments made in their frameworks or face monetary penalties.

These requirements would formalize practices that Anthropic and many other frontier AI companies already follow. At Anthropic, we publish our Responsible Scaling Policy, detailing how we evaluate and mitigate risks as our models become more capable. We release comprehensive system cards that document model capabilities and limitations. Other frontier labs (Google DeepMind, OpenAI, Microsoft) have adopted similar approaches while vigorously competing at the frontier. Now all covered models will be legally held to this standard. The bill also appropriately focuses on large companies developing the most powerful AI systems, while providing exemptions for startups and smaller companies that are less likely to develop powerful models and should not bear unnecessary regulatory burdens.

SB 53's transparency requirements will have an important impact on frontier AI safety. Without it, labs with increasingly powerful models could face growing incentives to dial back their own safety and disclosure programs in order to compete. But with SB 53, developers can compete while ensuring they remain transparent about AI capabilities that pose risks to public safety, creating a level playing field where disclosure is mandatory, not optional.

Looking ahead

SB 53 provides a strong regulatory foundation, but we can and should build upon this progress in the following areas and we look forward to working with policymakers to do so:

- The bill currently decides which AI systems to regulate based on how much computing power (FLOPS) was used to train them. The current threshold (10^26 FLOPS) is an acceptable starting point but there's always a risk that some powerful models may not be covered.
- Similarly, developers should be required to provide greater detail about the tests, evaluations, and mitigations they undertake. When we share our safety research, document our red team testing, and explain our deployment decisions—as we have done alongside industry players via the Frontier Model Forum—it strengthens rather than weakens our work.
- Lastly, regulations need to evolve as AI technology advances. Regulators should have the ability to update rules as needed to keep up with new developments and maintain the right balance between safety and innovation.

We commend Senator Wiener and Governor Newsom for their leadership on responsible AI governance. The question isn't whether we need AI governance—it's whether we'll develop it thoughtfully today or reactively tomorrow. SB 53 offers a solid path toward the former. We encourage California to pass it, and we look forward to working with policymakers in Washington and around the world to develop comprehensive approaches that protect public interests while maintaining America's AI leadership.

Related content

Donating the Model Context Protocol and establishing the Agentic AI Foundation

Read more →

Accenture and Anthropic launch multi-year partnership to move enterprises from AI pilots to production

Read more →

Snowflake and Anthropic announce $200 million partnership to bring agentic AI to global enterprises

Read more →

Products

Claude

Claude Code

Claude and Slack

Claude in Excel

Skills

Max plan

Team plan

Enterprise plan

Download app

Pricing

Log in to Claude

Models

Opus

06: Governor Newsom vetoed to return the bill without signature (Source: Official website)

OFFICE OF THE GOVERNOR

SEP 29 2024

To the Members of the California State Senate:

I am returning Senate Bill 1047 without my signature.

This bill would require developers of large artificial intelligence (AI) models, and those providing the computing power to train such models, to put certain safeguards and policies in place to prevent catastrophic harm. The bill would also establish the Board of Frontier Models – a state entity – to oversee the development of these models.

California is home to 32 of the world's 50 leading AI companies, pioneers in one of the most significant technological advances in modern history. We lead in this space because of our research and education institutions, our diverse and motivated workforce, and our free-spirited cultivation of intellectual freedom. As stewards and innovators of the future, I take seriously the responsibility to regulate this industry.

This year, the Legislature sent me several thoughtful proposals to regulate AI companies in response to current, rapidly evolving risks – including threats to our democratic process, the spread of misinformation and deepfakes, risks to online privacy, threats to critical infrastructure, and disruptions in the workforce. These bills, and actions by my Administration, are guided by principles of accountability, fairness, and transparency of AI systems and deployment of AI technology in California.

GOVERNOR GAVIN NEWSOM • SACRAMENTO, CA 95814 • (916) 445-2841

SB 1047 magnified the conversation about threats that could emerge from the deployment of AI. Key to the debate is whether the threshold for regulation should be based on the cost and number of computations needed to develop an AI model, or whether we should evaluate the system's actual risks regardless of these factors. This global discussion is occurring as the capabilities of AI continue to scale at an impressive pace. At the same time, the strategies and solutions for addressing the risk of catastrophic harm are rapidly evolving.

By focusing only on the most expensive and large-scale models, SB 1047 establishes a regulatory framework that could give the public a false sense of security about controlling this fast-moving technology. Smaller, specialized models may emerge as equally or even more dangerous than the models targeted by SB 1047 – at the potential expense of curtailing the very innovation that fuels advancement in favor of the public good.

Adaptability is critical as we race to regulate a technology still in its infancy. This will require a delicate balance. While well-intentioned, SB 1047 does not take into account whether an AI system is deployed in high-risk environments, involves critical decision-making or the use of sensitive data. Instead, the bill applies stringent standards to even the most basic functions — so long as a large system deploys it. I do not believe this is the best approach to protecting the public from real threats posed by the technology.

Let me be clear – I agree with the author – we cannot afford to wait for a major catastrophe to occur before taking action to protect the public. California will not abandon its responsibility. Safety protocols must be adopted. Proactive guardrails should be implemented, and severe consequences for bad actors must be clear and enforceable. I do not agree, however, that to keep the public safe, we must settle for a solution that is not informed by an empirical trajectory analysis of AI systems and capabilities. Ultimately, any framework for effectively regulating AI needs to keep pace with the technology itself.

To those who say there's no problem here to solve, or that California does not have a role in regulating potential national security implications of this technology, I disagree. A California-only approach may well be warranted – especially absent federal action by Congress – but it must be based on empirical evidence and science. The U.S. AI Safety Institute, under the National Institute of Science and Technology, is developing guidance on national security risks, informed by evidence-based approaches, to guard against demonstrable risks to public safety. Under an Executive Order I issued in September 2023, agencies within my Administration are performing risk analyses of the potential threats and vulnerabilities to California's critical infrastructure using AI. These are just a few examples of the many endeavors underway, led by experts, to inform policymakers on AI risk management practices that are rooted in science and fact. And endeavors like these have led to the introduction of over a dozen bills regulating specific, known risks posed by AI, that I have signed in the last 30 days.

I am committed to working with the Legislature, federal partners, technology experts, ethicists, and academia, to find the appropriate path forward, including legislation and regulation. Given the stakes – protecting against actual threats without unnecessarily thwarting the promise of this technology to advance the public good – we must get this right.

For these reasons, I cannot sign this bill.

Sincerely,

Gavin Newsom

07: Senator Wiener mentioning about Anthropics letter as one major feedback and explicitly calling them “World Leaders” (Source: Official website)

https://sd11.senate.ca.gov/news/senator-wieners-groundbreaking-artificial-intelligence-bill-advances-assembly-floor-amendments?

SCOTT WIENER

PRESS RELEASE

Senator Wiener’s Groundbreaking Artificial Intelligence Bill Advances To The Assembly Floor With Amendments Responding To Industry Engagement

AUGUST 15, 2024

SACRAMENTO – The Assembly Appropriations Committee passed Senator Scott Wiener’s (D-San Francisco) Senate Bill 1047 with significant amendments introduced by the author. SB 1047 is legislation to ensure the safe development of large-scale artificial intelligence systems by establishing clear, predictable, common-sense safety standards for developers of the largest and most powerful AI systems. The bill will now advance to the Assembly floor. It will be eligible for a vote on August 20th and must pass by August 31st.

“The Assembly will vote on a strong AI safety measure that has been revised in response to feedback from AI leaders in industry, academia, and the public sector,” said **Senator Wiener**. “We can advance both innovation and safety; the two are not mutually exclusive. While the amendments do not reflect 100% of the changes requested by Anthropic—a world leader on both innovation and safety—we accepted a number of very reasonable amendments proposed, and I believe we’ve addressed the core concerns expressed by Anthropic and many others in the industry. These amendments build on significant changes to SB 1047 I made previously to accommodate the unique needs of the open source community, which is an important source of innovation.

“With Congress gridlocked over AI regulation—aside from banning Tik Tok, Congress has passed no major technology regulation since

The major amendments to SB 1047, which will be published by the Senate in the coming days, are:

- **Removing perjury** – Replace criminal penalties for perjury with civil penalties. There are now no criminal penalties in the bill. Opponents had misrepresented this provision, and a civil penalty serves well as a deterrent against lying to the government.
- **Eliminating the FMD** – Remove the proposed new state regulatory body (formerly the Frontier Model Division, or FMD). SB 1047’s enforcement was always done through the AG’s office, and this amendment streamlines the regulatory structure without significantly impacting the ability to hold bad actors accountable. Some of the FMD’s functions have been moved to the existing Government Operations Agency.
- **Adjusting legal standards** - The legal standard under which developers must attest they have fulfilled their commitments under the bill has changed from “reasonable assurance” standard to a standard of “reasonable care,” which is defined under centuries of common law as the care a reasonable person would have taken. We lay out a few elements of reasonable care in AI development, including whether they consulted NIST standards in establishing their safety plans, and how their safety plan compares to other companies in the industry.
- **New threshold to protect startups’ ability to fine-tune open sourced models** – Established a threshold to determine which fine-tuned models are covered under SB 1047. Only models that were fine-tuned at a cost of at least $10 million are now covered. If a model is fine-tuned at a cost of less than $10 million dollars, the model is not covered and the developer doing the fine tuning has no obligations under the bill. The overwhelming majority of developers fine-tuning open sourced models will not be covered and therefore will have no obligations under the bill.
- **Narrowing, but not eliminating, pre-harm enforcement** – Cutting the AG’s ability to seek civil penalties unless a harm has occurred or there is an imminent threat to public safety.

SB 1047 is supported by both of the top two most cited AI researchers of all time: the “Godfathers of AI,” Geoffrey Hinton and Yoshua Bengio. Today, **Professor Bengio** published an op-ed in Fortune in support of the bill.

SCOTT WIENER

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.”

Inaccurate claims about the bill have spread online, leading to divided opinions among AI leaders.

In recent weeks, more AI industry leaders have come out in support of SB 1047. Simon Last, co-founder of Notion, was the latest to express support in an op-ed published last week.

Experts at the forefront of AI have expressed concern that failure to take appropriate precautions could have severe consequences, including risks to critical infrastructure, cyberattacks, and the creation of novel biological weapons. A recent survey found 70% of AI researchers believe safety should be prioritized in AI research more while 73% expressed “substantial” or “extreme” concern AI would fall into the hands of dangerous groups.

In line with President Biden’s Executive Order on Artificial Intelligence, and their own voluntary commitments, several frontier AI developers in California have taken great strides in pioneering safe development practices – implementing essential measures such as cybersecurity protections and safety evaluations of AI system capabilities.

Last September, Governor Newsom issued an Executive Order directing state agencies to begin preparing for AI and assess the impact of AI on vulnerable communities. The Administration released a report in November examining AI’s most beneficial uses and potential harms.

SB 1047 balances AI innovation with safety by:

- Setting clear standards for developers of AI models with computing power greater than 10^{26} floating-point operations that cost over $100 million to train and would be substantially more powerful than any AI in existence today

SCOTT WIENER

- Creating whistleblower protections for employees of frontier AI laboratories
- Empowering California’s Attorney General to take legal action in the event the developer of an extremely powerful AI model causes severe harm to Californians or if the developer’s negligence poses an imminent threat to public safety
- Establishing a new public cloud computer cluster, CalCompute, to enable startups, researchers, and community groups to participate in the development of large-scale AI systems and align its benefits with the values and needs of California communities

SB 1047 is coauthored by Senator Roth (D-Riverside), Senator Susan Rubio (D-Baldwin Park) and Senator Stern (D-Los Angeles) and sponsored by the Center for AI Safety Action Fund, Economic Security Action California, and Encode Justice.

###