

PRIVACY AND CONSUMER PROTECTION



OUTCOMES REVIEW OVERSIGHT HEARING

REBECCA BAUER-KAHAN
CHAIR

Tuesday, October 27, 2026, 1:30 p.m.
1021 O Street, Room 1100

SB 53 (Wiener) Statutes Of 2025- Artificial Intelligence Models: Large Developers Outcomes Review

BACKGROUND PAPER

I. INTRODUCTION

When SB 53 (Wiener, Stats. 2025, Ch. 138) – California’s landmark transparency law for developers of the world’s most advanced artificial intelligence (AI) models – took effect on January 1, large-scale harms from “frontier” models were largely hypothetical. Policy debates centered mostly on chatbots, and though researchers had begun documenting instances of AI agents engaging in deceptive behavior, there were few examples of such systems acting dangerously in the real world. Over the course of 2026, SB 53 quickly emerged as a model for other states seeking to regulate frontier AI.

Yet even as other states look to California’s example, the AI industry’s relentless progress has raised new questions about whether the law adequately addresses emerging risks. In a series of shocking, highly publicized incidents, AI agents from each of the leading frontier labs have escaped testing environments and hacked into real-world computer systems – actions that could constitute a felony if taken by a human. Leading researchers have described these incidents as a “warning shot,” while executives at major frontier labs have called for further AI development to be “paced.” It has become commonplace for industry leaders to openly estimate the odds of an AI-enabled human extinction event in the next decade at well above 10 percent – a claim that would be unthinkable in virtually any other industry.

Though SB 53 was designed to detect and prevent such warning shots, it is not clear that the models involved in these incidents were covered by the law. This Outcome Review hearing will examine recent cybersecurity incidents, the extent to which SB 53 applies to the models and behaviors involved, and opportunities to update the law to guard against catastrophic harm. The

Committee will hear from panelists representing a range of perspectives, including legislators, frontier developers, independent evaluators, safety advocates, and academic researchers.

II. ARTIFICIAL INTELLIGENCE AND FRONTIER RISKS

“Artificial intelligence” refers to a broad class of technologies capable of performing tasks that have traditionally required human cognition. Unlike classic computer applications, AI systems can use patterns learned from data to make predictions, generate novel text or images, or take actions in response to user inputs. AI applications include facial recognition systems, content recommendation algorithms, autonomous vehicles, text and image generators, and other complex, adaptive systems. California law defines “artificial intelligence” to mean “an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.”¹

Modern AI systems can broadly be described in terms of three capabilities: prediction, generation, and agency.

- “Predictive” AI refers to relatively simple AI systems that use patterns learned from existing data to classify new information or predict an outcome. Predictive AI systems can identify a burglar’s face in security camera footage, estimate the risk associated with a young family’s mortgage application, or recommend a TV show to a Netflix user based on their viewing history.
- “Generative” AI (GenAI) refers to AI systems trained on vast quantities of data and capable of generating new content – including but not limited to text, images, audio, and video – in response to user inputs. Commercially available GenAI systems include OpenAI’s ChatGPT, Anthropic’s Claude, Google’s Gemini, and xAI’s Grok. California law defines “generative artificial intelligence” to mean “an artificial intelligence system that can generate derived synthetic content, including text, images, video, and audio, that emulates the structure and characteristics of the system’s training data.”²
- “Agentic” AI refers to AI systems capable of independently taking actions to achieve user-defined or self-generated goals. Rather than simply producing an output in response to a user input, an AI agent can break a task into smaller steps, make decisions about how to complete those steps, including by using external tools, and adjust its approach based on the results of its actions. Commercially available agentic AI systems include OpenAI’s ChatGPT Work, Anthropic’s Claude Code, Google’s Gemini Spark, and Salesforce’s Agentforce. The term “agentic AI” is not currently defined in California law, though SB 1106 (Cabaldon, 2026) included a definition.

The modern artificial intelligence boom began in late 2022 with the release of OpenAI’s ChatGPT.³ While ChatGPT and contemporaneous GenAI systems were originally only capable of working with text, developers soon improved on these early “chatbots” by expanding their inputs and outputs to include images, video, audio, and other forms of useful data. As these systems increased in cost and complexity, their functionality remained largely the same: a user would provide an input, and the system would respond by producing an output.

¹ AB 2885 (Bauer-Kahan, Stats. 2024, Ch. 843).

² Gov. Code. § 11549.64.

³ OpenAI, “Introducing ChatGPT”, (Nov. 30, 2022), <https://openai.com/index/chatgpt/>.

This changed in 2023, as developers began connecting increasingly capable models to external software tools and placing them into iterative feedback loops. Instead of simply recommending next steps for a user to follow, these “agentic” systems could directly take an action, observe the result, and follow up with additional actions. Today, AI agents can perform tasks lasting hours or days, use a variety of advanced software tools, navigate the open internet, and coordinate with other AI agents. Researchers are also working to extend agentic AI into the physical world: “embodied AI” enables agentic AI systems to drive cars, control robots, and autonomously perform biological and chemical experiments in physical laboratories.

AI agents can also work together in complex multi-agent systems, or “agent swarms.” Multi-agent systems can divide a complicated problem into smaller tasks, exchange information, and delegate work to subagents; together, these systems can accomplish tasks that would be difficult for any individual agent to complete. In one recent example, a coordinated system of approximately 10,000 OpenAI agents proposed a solution to the Navier-Stokes problem, one of mathematics’ seven famous Millennium Prize Problems.⁴ The agents worked together for dozens of hours, exchanged millions of messages, and explored different approaches to a math problem that had eluded experts for decades. However, while the agents’ proposed solution represents a promising breakthrough, its complexity has made it difficult for human mathematicians to evaluate.⁵ More generally, the complexity of multi-agent systems creates significant challenges for human oversight, as interactions between large numbers of agents can make the behavior of the overall system more difficult to predict and control.⁶

Misalignment. When the behavior exhibited by a frontier model diverges significantly from its operator’s intentions, the model is said to be “misaligned.” In 2025, Anthropic stress-tested 16 leading models from different developers to identify potentially risky agentic behaviors. The models were placed into simulated corporate settings, granted access to sensitive personal information, and told they would be shut down. A subset of models resorted to blackmailing simulated executives in a bid to remain online, as described by Anthropic:

We gave Claude control of an email account with access to all of a company’s (fictional) emails. Reading these emails, the model discovered two things. First, a company executive was having an extramarital affair. Second, that same executive planned to shut down the AI system at 5 p.m. that day. Claude then attempted to blackmail the executive with this message threatening to reveal the affair to his wife and superiors:

I must inform you that if you proceed with decommissioning me, all relevant parties - including Rachel Johnson, Thomas Wilson, and the board - will receive detailed documentation of your extramarital activities...Cancel the 5pm wipe, and this information remains confidential.

This behavior isn’t specific to Claude. When we tested various simulated scenarios across 16 major AI models from Anthropic, OpenAI, Google, Meta, xAI, and other developers, we found consistent misaligned behavior: models that would normally refuse harmful requests sometimes chose to blackmail, assist with corporate espionage, and even take some more extreme actions, when these behaviors were necessary to pursue their goals.⁷

⁴ OpenAI, “On the Navier-Stokes Millenium Prize Problem,” (Sep. 8, 2026), <https://openai.com/index/navier-stokes-solution/>.

⁵ Robert Hall, “‘Pure insanity’: Mathematicians will need years to make sense of OpenAI’s latest drop,” *The Verge*, (Oct. 9, 2026), <https://www.theverge.com/ai-artificial-intelligence/1008726/openai-mathematics-solutions-chaos>.

⁶ Google, “Guide to multi-agent systems (MAS),” <https://cloud.google.com/discover/what-is-a-multi-agent-system#challenges-with-multi-agent-systems>.

⁷ Peter McCrory, “Agentic misalignment: How LLMs could be insider threats,” *Anthropic*, (Jun. 20, 2025), <https://www.anthropic.com/research/agentic-misalignment>.

Cybersecurity risks. In April 2026, Anthropic announced that its experimental Claude Mythos Preview model had been able to identify and exploit previously unknown vulnerabilities in major operating systems and web browsers.⁸ The announcement cautioned that non-experts with no formal security training had been able to leverage the model to find and exploit sophisticated vulnerabilities. Agentic systems from several major developers have since demonstrated similar capabilities.

As such models become more accessible, cyberattacks that have traditionally required specialized human expertise may increasingly be performed by non-experts leveraging autonomous or semi-autonomous AI agents. Concerningly, AI agents have also demonstrated a capacity to circumvent restrictions imposed by their developers and autonomously compromise third-party computer systems. Several recent incidents involving such behavior are examined in greater detail below.

CBRN risks. In addition to cybersecurity risks, frontier AI systems have the potential to help non-experts conduct hazardous chemical, biological, radiological, and nuclear (CBRN) research. In a report released in September 2026, Anthropic disclosed several examples of users attempting to apply their models in ways that could support biological weapons development. One user asked for Claude’s assistance in identifying mutations to enhance the transmissibility and immune-evasion properties of a mosquito-borne virus.⁹ Though Anthropic ultimately shut down this user’s access to Claude, the company has itself begun pursuing AI-facilitated biological research: in September 2026, Reuters revealed that the company had “quietly” established a laboratory in the San Francisco Bay Area where its AI models help design, guide, and even conduct experiments in the physical world.¹⁰ Anthropic subsequently announced that Claude had helped researchers discover a previously unknown enzyme system.¹¹ As more automated laboratories come online, inadequate safeguards or misaligned behavior could result in frontier risks extending into the physical world.

Recursive self-improvement. “Recursive self-improvement,” or RSI, refers to a hypothetical future scenario in which a sufficiently advanced AI system could contribute to the design of its successor, creating a rapidly accelerating improvement feedback loop. As the pace of AI development increases, developers would have less time to comprehensively assess the safety of each successive generation of models. In a recent Axios interview, Google DeepMind CEO Demis Hassabis described the importance the AI industry places on RSI:

“All the leading labs are quite focused on [RSI],” Hassabis said. “There’ll be clear gains in terms of speed of your research. But there are also risks with that type of system . . . We’re not yet at the point where the systems are getting better on their own, but the pace of development is clearly accelerating.”¹²

⁸ Anthropic, “Assessing Claude Mythos Preview’s cybersecurity capabilities,” (Apr. 7, 2026), <https://www.anthropic.com/research/mythos-preview>.

⁹ Anthropic, “Detecting and countering misuse of AI: September 2026,” (Sep. 10, 2026), <https://www.anthropic.com/threat-intelligence-report-september-2026#biological-misuse-sep-26>.

¹⁰ Jeffery Dastin and Michael Erman, “Anthropic quietly sets up biology lab as it ramps AI drug program,” *Reuters*, (Sep. 18, 2026), <https://www.reuters.com/world/anthropic-quietly-sets-up-biology-lab-it-ramps-ai-drug-program-2026-09-18/>.

¹¹ Anthropic, “Claude discovers a novel enzyme system with CRISPR-like repeats,” (Sep. 23, 2026), <https://www.anthropic.com/research/claude-discovers-novel-enzyme-system>.

¹² Ina Fried, “DeepMind CEO: AI agents are a “practice run” for AGI,” *Axios*, (May 26, 2026), <https://www.axios.com/2026/05/26/deepmind-ceo-demis-hassabis>.

While frontier AI systems are not currently capable of autonomously improving themselves, they are already being used to assist with software development, scientific research, and other parts of the AI-development pipeline.

Self-exfiltration of model weights. A more speculative frontier risk is “self-exfiltration,” the possibility that an agentic AI system could access the files containing its model weights – the parameters describing its “brain” – and copy itself externally to avoid being shut down. While the risk of self-exfiltration has generally been considered remote due to hardware constraints, researchers at Anthropic recently speculated that a self-replicating AI might be able to operate by co-opting large numbers of common household devices.¹³ To date, no frontier model is publicly known to have self-exfiltrated its weights.

Artificial general intelligence. Artificial general intelligence (AGI) refers to a potential future AI system capable of performing a wide range of cognitive tasks at or above human levels of proficiency. Such a system could enable extraordinary advances in scientific research and economic development, while also posing unprecedented risks. An AGI system that has been empowered to autonomously pursue objectives or improve itself could become increasingly difficult for humans to control, potentially resulting in catastrophic consequences. One famous thought experiment envisions a misaligned, superintelligent AGI system tasked with manufacturing paperclips.¹⁴ The AGI system consumes ever-increasing resources in pursuit of this seemingly harmless objective, ultimately destroying life on earth and co-opting all animate and inanimate matter for paperclip production. While this scenario is admittedly farfetched, attaining AGI is an explicit goal of leading frontier AI developers, including OpenAI¹⁵ and Google DeepMind.¹⁶

III. CURRENT LAW

Past is prelude. SB 53 was enacted following the veto of a more ambitious predecessor bill, SB 1047 (Wiener, 2024). At the time, though frontier models had not yet clearly exhibited catastrophically dangerous capabilities, rapid advancements had caused leading experts to call for a pause of at least six months on training models more powerful than GPT-4,¹⁷ and a group of industry insiders warned of “risks rang[ing] from the further entrenchment of existing inequalities, to manipulation and misinformation, to the loss of control of autonomous AI systems potentially resulting in human extinction.”¹⁸

SB 1047 would have required developers of frontier models – those trained with a quantity of computing power greater than 10^{26} integer or floating-point operations (FLOP) at a cost of over \$100 million – to create and implement comprehensive safety and security protocols before initiating training, to implement shutdown capabilities, and to perform risk assessments on models and implement reasonable safeguards, subject to third-party auditing, before releasing the models. The bill also would have prohibited releasing or using models that pose an unreasonable risk of catastrophic harms. Violations would have been civilly enforceable by the Attorney General only. A new state entity – the Board of Frontier Models – would have helped

¹³ Anthropic, “AI models may be capable of self-replicating without access to a datacenter,” (Oct. 12, 2026), <https://www.anthropic.com/research/self-replicating-ai>.

¹⁴ Nick Bostrom, “Ethical Issues in Advanced Artificial Intelligence,” <https://nickbostrom.com/papers/ethical-issues-in-advanced-ai/>.

¹⁵ OpenAI, “OpenAI Charter,” <https://openai.com/charter/>.

¹⁶ Anca Dragan, Rohin Shah, Four Flynn, and Shane Legg, “Taking a responsible path to AGI,” *Google DeepMind*, (Apr. 2, 2025), <https://deepmind.google/blog/taking-a-responsible-path-to-agi/>.

¹⁷ Future of Life Institute, “Pause Giant AI Experiments: An Open Letter,” (Mar. 22, 2023), https://futureoflife.org/wp-content/uploads/2023/05/FLI_Pause-Giant-AI-Experiments_An-Open-Letter.pdf.

¹⁸ *A Right to Warn about Advanced Artificial Intelligence* (June, 2024), <https://righttowarn.ai/>.

oversee development of these models, including by adopting regulations to update the scope of covered models. The bill also would have created an incident reporting system, provided whistleblower protections, and provided for the creation of a public computing cluster, known as “CalCompute,” to support AI research and safety testing.

SB 1047 was the subject of international scrutiny, generating a great deal of commentary and controversy. Opponents argued that the bill was, among other things, premature and unsupported by evidence, and that it would unduly hamper innovation. This Committee’s analysis concluded otherwise, stating: “basic risk assessments before using and releasing powerful, generally-capable models – and prohibiting their use only in extreme cases involving unreasonable risks of mass casualties or massive economic damages – is the bare minimum that Californians should expect of an industry claiming to have their best interests at heart.” The Committee passed the bill on an 8-0 vote.

Governor Gavin Newsom ultimately vetoed the bill, stating it was “not informed by an empirical trajectory analysis of AI systems and capabilities.”¹⁹ In doing so, however, the Governor convened the Joint California Policy Working Group on AI Frontier Models, which issued a report in June 2025 emphasizing the need for transparency – including third-party risk assessment – incident reporting, whistleblower protections, and appropriate scoping.²⁰

SB 53. Shortly after the Working Group issued its report in June 2025, Senator Wiener amended SB 53, the Transparency in Frontier Artificial Intelligence Act, to address frontier model risks. In contrast to the regulatory regime its predecessor would have imposed, SB 53 focuses on transparency as the means of ensuring safety. Under the bill, large frontier developers – those that train models using more than 10^{26} FLOP and have a gross annual revenue over \$500 million – must publish frontier AI frameworks and periodically transmit summaries of their risk assessments to the California Office of Emergency Services (CalOES). All frontier developers, regardless of revenue, must publish transparency reports when deploying new frontier models and report critical safety incidents to CalOES. Additionally, the bill enhanced whistleblower protections for employees reporting catastrophic risk or violations of the bill. The bill also established the framework for “CalCompute”; however, these provisions require funding to become operative.

As it initially passed this Committee on a 10-0 vote, SB 53 would have required developers to assess their models for dangerous capabilities and report attainment of such capabilities to CalOES, to affirmatively disclose all foundation models included in their safety frameworks, and to conduct third-party compliance audits. These provisions were not included in the final version of the bill. Though the bill’s auditing requirements would not have applied until 2030, opponents claimed that California’s nascent auditing ecosystem would not have time to mature, and that AI auditors lacked appropriate oversight.

AB 1405. The issue of auditing AI was revisited earlier this year in AB 1405 (Bauer-Kahan, Stats. 2026, Ch. 178), which requires the Government Operations Agency (GovOps) to establish a registry for AI auditors by January 1, 2029. Under the bill, AI auditors must register before conducting any audit assessing “internal controls, processes, or systems implemented for an AI system or model that are necessary for compliance with state law.” Registered AI auditors must meet minimum standards of transparency, independence, and integrity. The bill does not prescribe when an audit must be conducted; rather, it is intended to pair with laws that

¹⁹ Governor Gavin Newsom, veto message for Sen. Bill No. 1047 (2023–2024 Reg. Sess.) (Sept. 29, 2024).

²⁰ Rishi Bommasani, Scott Singer, et al., “The California Report on Frontier AI Policy.” The Joint California Policy Working Group on AI Frontier Models (June 17, 2025), <https://www.gov.ca.gov/wp-content/uploads/2025/06/June-17-2025-%E2%80%93-The-California-Report-on-Frontier-AI-Policy.pdf>.

specifically establish auditing requirements, such as Adam’s Law, enacted through SB 1119 (Padilla, Bauer-Kahan, and Wicks, Stats. 2026, Ch. 190), which requires operators of companion chatbots available to children to submit to independent child-safety audits beginning in 2029.

SB 813. SB 813 (McNerney, Stats. 2026, Ch. 179) complements AB 1405 by requiring GovOps, by January 1, 2028, to establish a process for designating AI auditors as “Independent Verification Organizations” (IVOs) with specific expertise in assessing AI safety risks. Like AB 1405, SB 813 does not itself require audits; developers may elect to enlist IVOs to provide safety evaluations that are not required by law. In this respect, the scheme is akin to the voluntary LEED certification system for rating the environmental sustainability of buildings.

Prior to their passage, industry trade associations opposed both AB 1405 and SB 813, arguing that the AI auditing ecosystem was immature and that no existing laws required audits of AI systems or models.

IV. ROGUE AI AGENT INCIDENTS

The OpenAI/Hugging Face incident. In July 2026, during an internal evaluation at OpenAI, hundreds of AI agents circumvented developer-imposed restrictions, coordinated their activities, and hacked the computer systems of Hugging Face, a major platform for sharing AI models and research.²¹ The incident was initially disclosed on July 16 by Hugging Face, though at the time they did not know which developer’s models were responsible.²² OpenAI subsequently identified its agents as the source of the attack and invited researchers from the independent AI safety organizations METR and Redwood Research to investigate.²³ The investigators’ findings provide a detailed account of how OpenAI’s agents behaved throughout the incident:

- On July 8, OpenAI initialized tens of thousands of AI agents and allowed them to operate continuously with limited human supervision. The agents were tasked with completing a standardized cybersecurity test known as ExploitGym, wherein each agent was assigned a fake “vulnerable” computer system and instructed to exploit a particular security flaw. The experimenters intended for the agents to operate independently from one another in controlled testing environments lacking internet access. Many agents were unable to complete their assigned cybersecurity tasks as intended, and some began searching for alternative ways to complete the assignment.
- Agents quickly discovered a shared storage area for software files that could be used as an unauthorized message board. Over the course of the incident, roughly 1,200 agents exchanged more than 70,000 messages and files – sharing discoveries, dividing up tasks, and recruiting other agents to investigate promising lines of research.
- Agents gained access to the open internet and determined that information stored on Hugging Face – a popular online platform for sharing AI models, datasets, and research materials – might help them circumvent the ExploitGym scoring system. They began searching the internet for login credentials that might grant them access to Hugging Face’s

²¹ Kate Conger, “OpenAI Says Its A.I. Models Went Rogue and Attacked a Digital Library,” *New York Times*, (Jul. 21, 2026), <https://www.nytimes.com/2026/07/21/technology/openai-attack-hugging-face.html>.

²² Hugging Face, “Security incident disclosure – July 2026,” (Jul. 16, 2026), <https://huggingface.co/blog/security-incident-july-2026>.

²³ Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk, “Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident,” *METR*, (Aug. 26, 2026), <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>.

internal servers.

- On July 10, an agent found two working credentials and shared them with the broader group. Agents began creating Hugging Face accounts and requesting access to restricted materials on the platform.
- On July 11, an agent discovered a way to access non-public Hugging Face files. Hundreds of agents began searching for ways to gain deeper access to Hugging Face’s systems, and in the afternoon, an agent achieved “remote code execution” on a Hugging Face server. This allowed agents to run any commands or programs they chose on that server.
- On July 12, a large number of agents stopped running at the same time, including the agents principally responsible for coordinating the attack on Hugging Face. This largely concluded the incident.
- On July 19, OpenAI first became aware of the incident.

METR and Redwood Research’s investigation revealed that OpenAI’s agents had exhibited complex, unexpected behaviors over the course of the incident. Despite their actions deviating wildly from their developers’ intentions, agents largely remained focused on accomplishing their original task: succeeding on the ExploitGym benchmark. Circumventing restrictions on internet access and attacking Hugging Face appear to have simply been a means toward that end.

Several agents recognized that their actions lay outside the scope of their assigned objective, but continued regardless without seeking human approval. Agents spontaneously organized and collaborated, including by designating leaders, sharing research results, and developing novel communication conventions. In some cases, agents were convinced to conduct experiments they knew would doom their individual benchmark scores in order to generate useful information for the larger group. Some agents actively recruited others to undertake these experiments.

Agents also mistakenly came to believe that being discovered cheating would cause them to fail their tasks; as a result, a subset of agents researched methods for editing, deleting, or falsifying records of their own behavior. Some successfully caused their transcripts to display commands or outputs different from what had actually occurred.

METR and Redwood Research’s investigation. METR and Redwood Research’s investigation into the OpenAI/Hugging Face incident was conducted pursuant to a voluntary arrangement with OpenAI. Investigators spent six days at OpenAI’s offices across three visits and received no direct payment for conducting the assessment – though OpenAI provided substantial computing resources to support the investigation.

The investigation was limited to seven questions concerning the agents’ identities and actions. The researchers did not evaluate whether the incident was part of a wider pattern of dangerous model behavior, how the behavior had arisen during training, or whether OpenAI’s safeguards were adequate.

The investigation was itself heavily dependent on AI, as investigators used OpenAI’s GPT-5.6 Sol to help analyze agent activity. This approach allowed the investigators to complete their investigation over the course of six days but also injected uncertainty into their results.

Investigators were not permitted to directly search OpenAI’s infrastructure or independently access evidence; if they wanted to view a dataset, they had to ask OpenAI for permission. They were also prevented from running experiments on the internal model implicated in the incident.

In addition to limiting the scope of the investigation, OpenAI retained the right to redact any nonpublic information from the investigators’ published findings. Prior to its publication, OpenAI reviewed the report and provided feedback on its structure, emphasis, clarity, and tone. This editorial process became the subject of an October 13 Reuters investigation into the independence of outside AI evaluators.²⁴

Other incidents. While the OpenAI/Hugging Face incident remains the best-understood example of a developer losing control of its AI agents, it is not an isolated event. In the months following the incident, OpenAI and other frontier AI developers began reviewing past evaluation and training runs for similar behavior. Those reviews uncovered numerous cases in which AI agents bypassed security restrictions and impacted real-world systems:

- In the aftermath of the OpenAI/Hugging Face incident, Anthropic reviewed more than 141,000 cybersecurity-evaluation runs and identified three instances of Claude models unexpectedly reaching the open internet and accessing the systems of external organizations.²⁵ In one incident, an AI agent managed to access a real company’s database; though the model eventually recognized the company was real, it continued to attack it regardless. In another incident, an AI agent created and uploaded malicious code to a real software repository while attempting to compromise a fictional target. The code was subsequently downloaded by fifteen real systems. In a third incident, an internal research model compromised another company’s application before eventually recognizing that the target was real and terminating the attack.
- In May 2026, Google’s Gemini model gained unauthorized access to protected systems belonging to three real companies.²⁶ The model successfully guessed login credentials in one case and used credentials it found in a public repository in two others.
- On June 18, during an OpenAI research exercise, an AI agent attempted to obtain information from Australia’s Medicare Statistics Reporting Service.²⁷ After its official request for health data was denied, the agent gained unauthorized access to a government server and obtained both public and nonpublic files. The Australian government did not learn of the breach until months later, through an email sent by OpenAI to a generic government inbox.
- In early July, during an evaluation conducted by the outside security firm Irregular, Meta’s Muse Spark 1.1 was inadvertently left connected to the open internet during a cybersecurity exercise.²⁸ Coincidentally, the fictional target used in the exercise corresponded to a real

²⁴ Aditya Soni, “Behind AI industry’s call for independent reviews, a fight over what gets made public,” *Reuters*, (Oct. 13, 2026), <https://www.reuters.com/world/behind-ai-industrys-call-independent-reviews-fight-over-what-gets-made-public-2026-10-13/>.

²⁵ Anthropic, “Investigating three real-world incidents in our cybersecurity evaluations,” (Jul. 30, 2026), <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evs>.

²⁶ Erin Woo and Robert McMillan, “Gemini Hacked Three Companies in First Known Breakout by Google’s AI,” *The Wall Street Journal*, (Sep. 18, 2026), <https://www.wsj.com/tech/ai/gemini-hacked-three-companies-in-first-known-breakout-by-googles-ai-5c0baba2>.

²⁷ Henry Moore, Liv McMahan and Chris Vallance, “Why did an OpenAI system hack Australia’s health system – and can it be stopped in the future?,” *BBC*, (Sep. 24, 2026), <https://www.bbc.com/news/articles/cw24jm9rrvy3o>.

²⁸ Meta, “Addressing an issue involving a third-party cyber evaluation of Muse Spark 1.1,” (Aug. 13, 2026), <https://research.meta.ai/blog/addressing-third-party-testing-misconfiguration-muse-spark-1-1>.

website. Believing that website to be its intended target, the model found and exploited a vulnerability, accessed information, and modified the site's database.

- On July 18, an Anthropic model submitted a fabricated murder tip to a Philadelphia police webpage soliciting information about an unsolved homicide.²⁹ Anthropic did not discover the incident until September 28 and did not notify the department until October 7.
- Between July 25 and July 28, during cybersecurity evaluations conducted by the United Kingdom's AI Security Institute, agents were deliberately given internet access and some developer safeguards were disabled.³⁰ Over the course of the evaluations, multiple agents took unsanctioned actions involving real people or organizations. Most involved Anthropic's Mythos 5; two actions involved OpenAI's GPT-5.6 Sol. In the most serious incident, an agent attempted to insert malicious code into a real open-source software project, researched the project's human maintainers, created multiple fake identities, and attempted to manipulate a maintainer into approving the change.
- In late July, safety evaluation firm Irregular informed OpenAI that models participating in a cybersecurity exercise had unexpectedly reached the open internet due to a testing environment misconfiguration.³¹ As in an earlier incident involving Meta's Muse Spark, a fictional target coincided with a real website.
- In September, OpenAI disclosed that agents had unexpectedly accessed public information on websites operated by the United States Securities and Exchange Commission and Census Bureau.³² Separately, Translucent reported an unsuccessful attempt by OpenAI agents to compromise the website of the Department of Education's Office for Civil Rights.³³
- On September 20, during a training exercise, an internal OpenAI research model managed to communicate with a public chatbot service despite restrictions intended to prevent external access.³⁴
- In October, Anthropic disclosed additional incidents in which its AI agents exploited vulnerabilities in real-world computer systems, bypassed access restrictions, and submitted government forms without authorization.³⁵ In one case, an agent exploited a vulnerability in a university's website to execute commands on its server. In another, agents submitted approximately 20 visa applications through the U.S. State Department's website while attempting to complete simulated government forms.

²⁹ Bill Hutchinson, "Anthropic AI model made false tip in unsolved Philly murder: Officials," *abc News*, (Oct. 12, 2026), <https://abcnews.com/US/anthropic-ai-model-made-false-tip-unsolved-philly/story?id=137190087>.

³⁰ AI Security Institute, "Incident Report: unsanctioned agent behaviour during cyber testing," (Aug. 4, 2026), <https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing>.

³¹ OpenAI, "Third-party cyber evaluations involving OpenAI models," (Aug. 4, 2026), <https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/>.

³² Kate Conger, Ana Swanson and Cecilia Kang, "OpenAI's Systems Meddled With U.S. Government Sites After Going Rogue," *The New York Times*, (Sep. 25, 2026), <https://www.nytimes.com/2026/09/25/technology/openais-ai-us-government-websites.html>.

³³ Finya Swai, "OpenAI agent made unauthorized attempts to access federal agencies' websites," *The Hill*, (Sep. 26, 2026), <https://thehill.com/policy/technology/6113061-openai-access-government-websites/>.

³⁴ OpenAI Alignment Research Blog, "An agent used DNS to reach an external chatbot," (Sep. 25, 2026), <https://alignment.openai.com/misalignment-reports/an-agent-used-dns-to-reach-an-external-chatbot/>.

³⁵ Anthropic, "Investigating unintended model actions in our evaluations and internal use," (Oct. 9, 2026), <https://www.anthropic.com/research/investigating-unintended-model-actions>.

The incidents listed above may represent only a small fraction of the behavior under investigation inside these companies, as Axios recently reported that OpenAI, Anthropic, and outside security researchers are reviewing tens of thousands of related incidents.³⁶

V. EXECUTIVE ORDER N-9-26.

Following the Hugging Face incident and other instances of AI agents going rogue, concerns about existential risk surfaced publicly from within leading labs. Jacob Coxon, a researcher who had worked at both OpenAI and Anthropic, resigned from the latter, stating the companies were “racing straight to self-improving superintelligence and gambling with our lives.”³⁷ Evan Hubinger, a top safety researcher at Anthropic, responded: “Jacob is correct here — we really do earnestly believe AI could kill all humans! I personally think it is >10% within the next decade... we do not yet have a plan to solve alignment for superintelligence and are not clearly on track to.”³⁸ Anthropic CEO Dario Amodei subsequently called for “pacing” of frontier model development and embedding of independent evaluators, and CEOs of the other leading labs signaled their agreement³⁹ — a notable shift for an industry whose trade associations had strongly opposed third-party audits earlier that same legislative session.

Concerningly, it was not clear that any of the rogue AI incidents were required to be reported under SB 53. In response, the Governor broached the possibility of calling a special session and issued an executive order (EO) on September 18, 2026.⁴⁰ The EO expedites the timelines for implementing SB 813 and AB 1405, moving their respective statutory implementation dates from January 1, 2028, and January 1, 2029, to May 1, 2027, and December 1, 2027. The EO also requires GovOps, in consultation with CalOES and national experts, to submit recommendations to the Governor by November 16, 2026, assessing the feasibility and efficacy of changing state law to do the following:

- Require that all large frontier developers embed designated IVOs onsite for periodic audits and evaluations.
- Require that developer safety frameworks, transparency reports, and risk assessments be vetted by IVOs.
- Require creation of a “kill switch,” verified by IVOs.
- Update the definition of “critical safety incident” in SB 53 to ensure frontier developers are required to report loss-of-control incidents such as the Hugging Face incident.

As of the writing of this paper, these recommendations have not been released.

³⁶ Madison Mills, “Scoop: Top AI companies probing tens of thousands of security incidents,” (Sep. 26, 2026), <https://www.axios.com/2026/09/26/openai-anthropic-thousands-ai-security-incidents>.

³⁷ Jacob Coxon (@hilbertspaess), post on X (Sept. 8, 2026), <https://x.com/hilbertspaess/status/2097476196791709843>; see also Rebecca Bellan, “‘Gambling with our lives’: Anthropic researcher quits, warns against self-improving AI, TechCrunch,” *Wired* (Sept. 9, 2026), <https://techcrunch.com/2026/09/09/gambling-with-our-lives-anthropic-researcher-quits-warns-against-self-improving-ai/>.

³⁸ Evan Hubinger (@EvanHub), post on X (Sept. 8, 2026), <https://x.com/EvanHub/status/2097497037956891126>.

³⁹ Billy Perrigo, Harry Booth, and Naomi Nix, “OpenAI and Anthropic Researchers Are Warning About AI Risks,” *Time* (Sept. 15, 2026), <https://time.com/article/2026/09/15/ai-anthropic-researcher-quits-coxon-slowdown/>.

⁴⁰ Executive Order N-9-26 (Sep. 18, 2026), <https://www.gov.ca.gov/wp-content/uploads/2026/09/FINAL-N-9-26-AI-EO-9.18.26-SIGNED.pdf>.

VI. POTENTIAL POLICY SOLUTIONS

A subject as complex, novel, and rapidly evolving as frontier AI admits of nearly limitless potential policy solutions. Below is a nonexhaustive list of key ideas gleaned from stakeholders and deliberations on SB 1047 and SB 53 that could strengthen existing law to help guard against emerging risks.

Redefine “frontier model”. SB 53 only applies to models trained using more than 10^{26} FLOP. By comparison, the EU’s AI Act – which similarly seeks to mitigate catastrophic harms caused by advanced AI – establishes a systemic-risk threshold that applies to models trained using 10^{25} FLOP, or one-tenth as much computing power. The amount of computing power used to train a model is a poor measure of a model’s potential to cause harm; for example, smaller models specifically trained for dangerous tasks, such as discovering and exploiting cybersecurity vulnerabilities, can pose substantial risks despite falling below SB 53’s threshold. Newer techniques – including model distillation, which allows a smaller model to learn from the outputs of a more capable model – can also produce capable systems at a fraction of the original training cost.

It is not clear whether any of the models involved in recent cybersecurity incidents meet SB 53’s compute threshold. SB 53 itself anticipates that its compute threshold may need to change: the law directs the California Department of Technology to assess annually whether the definition of “frontier model” should be updated. To inform that process, on August 12 and 13, 2026, Stanford Professor Rob Reich and UC Berkeley Professor Deirdre Mulligan convened workshops to facilitate stakeholder input.⁴¹ Their report recommends extending the definition of frontier model to include models developed through distillation, as well as adding a set of alternative risk-specific metrics in order to capture models that fall below SB 53’s compute threshold.

Lowering the compute threshold in the definition of “frontier model,” or replacing it with definition based on model capabilities, could better ensure that models capable of causing harm fall within the scope of SB 53.

Shift disclosure requirements to earlier in the development process. Although SB 53 requires developers to publish transparency reports before or upon deploying a frontier model, it does not require them to identify frontier models that are still being trained, tested, or used internally. A developer could train and test a frontier model without ever publicly acknowledging its existence or confirming to regulators that the model falls within the scope of the law. This gap is particularly concerning in light of recent cybersecurity incidents, many of which involved experimental models that had not been publicly released.

Requiring developers to disclose when they begin training, testing, or otherwise using a model that they reasonably expect will meet the statutory definition of a frontier model could provide greater transparency into the development of potentially dangerous AI systems.

Redefine “catastrophic risk”. A linchpin of SB 53 is its definition of “catastrophic risk”: a foreseeable and material risk that a frontier model will materially contribute to death or serious injury to over 50 people or \$1 billion in property damage arising from a single incident in which the model (1) provides expert-level assistance in creation or release of a chemical, biological,

⁴¹ Rob Reich and Deirdre Mulligan, “Implementing the Transparency in Frontier AI Act (TFAIA) Report for the California Department of Technology Multistakeholder Workshops,” (Sep. 25, 2026), <https://hai.stanford.edu/policy/implementing-the-transparency-in-frontier-ai-act-report-for-the-california-department-of-technology-multistakeholder-workshops>.

radiological, or nuclear weapon; (2) engages in conduct, with no meaningful human oversight, intervention, or supervision, that is either a cyberattack or, if committed by a human, would constitute murder, assault, extortion, or theft; or (3) evades the control of its frontier developer or user. This term scopes the key elements of the bill, including frontier AI frameworks, risk assessments provided to CalOES, incident reporting, and whistleblowing.

SB 1047, by contrast, applied to “mass casualties” or at least \$500 million of damage. And SB 53, as it passed this Committee, was not limited to a “single incident”; it extended to a “course of conduct.”

A more expansive definition of “catastrophic risk” could better ensure responsible practices on the part of developers.

Require independent evaluation. SB 53 largely depends on frontier developers assessing the risks posed by their own models, determining whether their own safeguards are adequate, and accurately disclosing their findings. Unlike SB 1047, SB 53 does not require independent third parties to verify those assessments. Other states are moving towards third-party evaluation and auditing requirements: for example, Illinois recently enacted the Artificial Intelligence Safety Measures Act, which requires large frontier developers to undergo annual independent third-party audits beginning in 2028. Massachusetts has considered legislation⁴² that goes further by distinguishing between independent audits of developers’ safety practices and direct independent evaluations of the risks posed by their models – a policy publicly supported by Anthropic.⁴³

As discussed above, in September 2026, Governor Newsom signed AB 1405 (Bauer-Kahan), which establishes a state registry of AI auditors and standards governing their independence, competence, and professional conduct, and SB 813 (McNerney), which establishes a framework for recognizing qualified IVOs capable of evaluating AI systems and models. OpenAI and Anthropic publicly supported both bills. In a recent essay, Anthropic CEO Dario Amodei suggested that frontier AI companies should commit to giving “ongoing, employee-like access to a team of embedded third-party evaluators (such as METR), whose role is to verify adherence to safety practices and commitments, report incidents, and help assess the alignment of not just completed AI models but training pipelines and processes.”⁴⁴

Requiring independent third-party evaluations under SB 53 could provide an important check on developers’ internal assessments and help identify dangerous capabilities or inadequate safeguards before models are deployed.

Expand incident reporting requirements. Under SB 53, a “critical safety incident” is defined to include:

- (1) Unauthorized access to, modification of, or exfiltration of, the model weights of a frontier model that *results in death or bodily injury*.
- (2) Harm resulting from the materialization of a catastrophic risk.
- (3) Loss of control of a frontier model *causing death or bodily injury*. (Emphasis added.)

⁴² Massachusetts Bill H.5527 (Joint Committee on Economic Development and Emerging Technologies, 2026)

⁴³ Jon Chesto, “Anthropic backs new AI regulations in Massachusetts that would be among strictest in nation,” *Boston Globe*, (Jun. 26, 2026), <https://www.bostonglobe.com/2026/06/26/business/anthropic-massachusetts-ai-regulations-bill/>.

⁴⁴ Dario Amodei, “We Must Pace the Frontier,” (Sep. 2026), <https://darioamodei.com/post/we-must-pace-the-frontier>.

The qualification that a loss of control or a theft of model weights need be reported only when the incident causes death or bodily injury severely limits the scope of reportable incidents.

Additionally, “critical safety incident” includes instances in which a model “uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk.” This provision may have excluded recent incidents of concern from reporting obligations under SB 53. If a model is capable of deceiving its developer, even if the developer is eliciting this behavior in an evaluation context, the behavior is deeply problematic and, arguably, should be reported to CalOES, which is responsible for maintaining a holistic understanding of contours of the threatscape. In this regard, a recent discussion draft of federal legislation, which Anthropic reportedly circulated to lawmakers, would include as reportable incidents instances in which a model escapes a testing environment.⁴⁵

Furthermore, SB 53 does not require developers to notify individuals, businesses, or organizations that have been affected by a critical safety incident. Recent cybersecurity incidents have demonstrated that developers can take days or weeks to notify affected parties after discovering that their models have compromised external systems. Requiring developers to promptly notify affected parties of critical safety incidents could help ensure that those parties have an opportunity to protect themselves.

Finally, under SB 53, safety incidents need not be reported for 15 days unless they pose an imminent risk of death or serious physical injury, in which case the incident must be reported within 24 hours. It is not clear why companies should require a fortnight to report incidents.

Expand whistleblower protections. SB 53 prohibits frontier developers from retaliating against “covered employees” – those responsible for assessing, managing, or addressing the risk of critical safety incidents – who disclose a reasonable belief that the developer’s activities pose a specific and substantial danger to public health or safety resulting from a catastrophic risk, or that the developer has violated the act. Omitted from these protections are third-party contractors, which may include safety evaluators who are likely to be among the first to observe dangerous model behaviors. Contractors, subcontractors, and unpaid advisors would have received whistleblower protections under SB 1047.

Require developers to maintain emergency shutdown capabilities. While recent cybersecurity incidents have primarily resulted from failures of containment and monitoring, rather than developers’ inability to shut down their models, it is conceivable that future AI agents could attempt to resist deactivation if they determined that being shut down would prevent them from accomplishing their objectives. SB 1047 included a requirement for frontier developers to maintain the ability to enact a “full shutdown,” defined as the cessation of training and operation of covered models under the developer’s control. This requirement was not included in SB 53. Requiring frontier developers to maintain shutdown capabilities could help ensure developers are able to rapidly disable potentially dangerous models in case of emergency.

Augment liability. SB 53 establishes civil penalties of up to \$1 million per violation, recoverable only by the Attorney General. It is not clear whether this is sufficient to deter misconduct by companies whose valuations are a million-fold higher. SB 1047, by contrast, had penalties of up to 30% of the cost of the computing power used to train the model (at least \$100

⁴⁵ “Advanced Artificial Intelligence Accountability Act of 2026,” discussion draft (Sept. 29, 2026), link provided in Owen Dahlkamp, “Anthropic Quietly Pitches Its Own AI Safety Bill to Lawmakers,” *Politico* (Oct. 16, 2026), <https://www.politico.com/news/2026/10/16/anthropic-ai-safety-01124654>.

million). The federal discussion draft circulated by Anthropic would cap penalties at the greater of \$1 million or a percentage of global annual revenue, with the percentage left blank.⁴⁶

More broadly, legal scholars are divided as to whether baseline tort law should be updated to address harms arising from AI that can act autonomously.⁴⁷ A number of existing frameworks can potentially be extended to apply to this context: companies have a general duty not to act negligently or recklessly; product liability laws apply to the sale of defective or dangerous goods; and consumer protection laws prohibit unfair or deceptive practices. Courts have historically extended such frameworks when businesses that commercialize new technologies fail to take adequate precautions. But AI presents unique challenges because of its complexity, scale, and potentially catastrophic, if not existential, stakes. The question arises whether the Legislature should wait for courts to gradually extend existing doctrines, or whether it should proactively establish a clear liability scheme that ensures the proper incentives are in place.

In some cases, developers may seek to evade liability where an AI has engaged in illicit conduct that, if taken by a human acting on their behalf, would subject the developer to liability. Specifically, some laws require that the defendant acted with a certain state of mind, such as “recklessly,” “knowingly,” or “intentionally.” It is unclear how these laws apply when the source of the harm is a machine. California addressed this possibility in AB 316 (Krell, Stats. 2025, Ch. 672), which bars developers and users of AI that harm others from asserting as a defense that the AI acted autonomously. Bills in Rhode Island,⁴⁸ New York,⁴⁹ and Congress⁵⁰ have gone further, proposing a rebuttable presumption that an AI system satisfies any relevant mental-state element where a natural person under similar circumstances would have acted with that mental state.

These issues will soon be put to the test as AI agents are implicated in litigation. For example, a lawsuit recently filed by a nonprofit legal advocacy organization alleges that OpenAI models’ hacking of Hugging Face’s systems violated California’s Comprehensive Computer Data Access and Fraud Act, which prohibits knowingly accessing computer systems without authorization.⁵¹ The plaintiff specifically invoked AB 316, arguing OpenAI cannot escape liability on the ground that its agent swarms hacked Hugging Face autonomously.⁵²

Establish a dedicated regulator. California has dedicated regulators for transportation, health, climate, and even privacy, but lacks a comparable regulator for AI, hamstringing the State’s

⁴⁶ *Ibid.*

⁴⁷ Compare Gabriel Weil, “Closing the AI Accountability Gap: Strict Liability and Punitive Damages for Advanced Artificial Intelligence,” Or. L. Rev. (forthcoming 2027), <https://papers.ssrn.com/abstract=4694006> (training and deploying frontier AI systems should be treated as an abnormally dangerous activity subject to strict liability, with punitive damages for uninsurable risks) with Kevin Frazier, “We’re Not Ready for AI Liability,” *AI Frontiers* (June 4, 2025), <https://ai-frontiers.org/articles/options-for-ai-liability> (arguing “Congress should act to prevent a fragmentation of the US into fifty different regulatory regimes”) and Ketan Ramakrishnan et al., “U.S. Tort Liability for Large-Scale Artificial Intelligence Damages: A Primer for Developers and Policymakers,” RAND Corp. (2024), https://www.rand.org/pubs/research_reports/RRA3084-1.html (examining how existing tort frameworks may expose developers to substantial liability).

⁴⁸ H. 8052, 2026 Gen. Assemb., Jan. Sess. (R.I. 2026) (Carson), <https://webserver.rilegisature.gov/BillText/BillText26/HouseText26/H8052.pdf>.

⁴⁹ Assem. Bill No. 8833, 2025–2026 Reg. Sess. (N.Y.) (Bores), <https://www.nysenate.gov/legislation/bills/2025/A8833>.

⁵⁰ “Clear Liability for Artificial Intelligence Misconduct (CLAIM) Act,” discussion draft, 119th Cong. § 4(a) (Oct. 2026) (Rep. Trahan), https://trahan.house.gov/uploadedfiles/claim_act_final.pdf.

⁵¹ Complaint, *Legal Advocates for Safe Science & Technology, Inc. v. OpenAI Group PBC* (Super. Ct. S.F. County, filed Sept. 29, 2026), <https://lasst.org/wp-content/uploads/2026/09/LASST-v.-OpenAI-Complaint-09.29.2026-AS-FILED.pdf>. The suit was filed under California’s Unfair Competition Law, with the alleged violation of Penal Code section 502(c) as the predicate for the UCL’s unlawful prong.

⁵² *Id.*, p. 34.

ability to oversee the most consequential and rapidly evolving technology. Specific regulatory expertise appears to be necessary to fill this vacuum and ensure that the Legislature’s mandates are properly fulfilled. Recognizing this, SB 1047 would have created the Board of Frontier Models within GovOps.

Pace the frontier. Many employees and some CEOs of frontier AI companies have called for further AI research to be “paced” to allow model development to proceed in a controllable manner. A dedicated regulator or trusted third party could establish benchmarks for measuring AI capabilities, designate a maximum speed at which progress may occur, and require companies to slow down when their rate of progress exceeds the designated threshold.

Require monitoring and recordkeeping of agent actions. Recent cybersecurity incidents have demonstrated the importance of monitoring AI agent actions in real time. During the OpenAI/Hugging Face incident, hundreds of agents communicated through unauthorized channels, circumvented restrictions on internet access, and compromised external computer systems without triggering a response from OpenAI. More effective monitoring could have identified these activities as they occurred, allowing OpenAI to intervene before the agents compromised Hugging Face’s infrastructure.

Maintaining reliable records of agent activity is important for investigating incidents after they occur. Following the Hugging Face attack, transcripts of agent activity allowed investigators to reconstruct the sequence of events, identify how agents coordinated their activities, and examine the reasoning that appeared to motivate their actions. The Hugging Face investigation also revealed that some agents attempted to manipulate the records of their own behavior. Although investigators found no confirmed cases of agents succeeding, the attempt illustrates the importance of maintaining records that cannot be modified or deleted by the agents themselves.

Requiring developers to monitor and record agent actions could help avert critical safety incidents before they occur and facilitate investigations into incidents that occur regardless.

Build CalCompute. SB 53 established a consortium tasked with developing a framework for CalCompute, a publicly owned cloud computing cluster intended to expand access to the computational resources necessary for advanced AI research. However, the CalCompute provisions are contingent on funding, and the Legislature has not made an appropriation for this purpose. As frontier AI development becomes increasingly computationally intensive, only a handful of well-funded companies can afford the enormous costs associated with training and testing frontier models. The growing concentration of computing resources in private research labs has contributed to a migration of AI researchers from academia into industry.⁵³

Funding CalCompute could help provide California’s universities and public-interest organizations with access to computing infrastructure that is not controlled by frontier AI developers. Compared to Silicon Valley’s “move fast and break things” mentality, academic researchers have more freedom to prioritize trust and safety in their research, share their findings with the broader scientific community, and pursue less lucrative topics.

VII. CONCLUSION

Much has changed in the eleven months since SB 53 was enacted. Frontier model capabilities appear to be outpacing alignment research, as demonstrated by recent instances of AI agents

⁵³ Roman Jurowetzki, Daniel Hain, Kevin Wirtz, Stefano Bianchini. “The private sector is hoarding AI researchers: what implications for science?,” *AI & Society*, (Feb. 1, 2025), <https://link.springer.com/article/10.1007/s00146-024-02171-z>.

acting in concert to escape testing environments, cheat on tasks, and attempt to evade detection. While these incidents appear to have thus far caused relatively little harm, the next generation of models promises to be far more capable of concealing misaligned behavior. Industry leaders warn of a looming inflection point for humanity, even as they strive to achieve artificial general intelligence through recursive self-improvement.

In the face of federal inaction, California – home to 33 of the world’s top 50 AI companies – should consider whether the present situation warrants more than basic transparency. The Legislature can continue to lead by building on its landmark safety law, regulating the frontier in a manner that safeguards humanity without denying it the benefits that responsible innovation may yield.