Understanding the Contours of the AI Threatscape

An Informational Hearing of the Senate Committee on Privacy, Digital Technologies, and Consumer Protection

By the Staff of the Senate Committee on Privacy, Digital Technologies, and Consumer Protection

---

I. Introduction

The proliferation of modern artificial intelligence (AI) has sparked significant debate about its implications across nearly every facet of life. This technology has the potential to revolutionize California and the world; however, it also raises reasonable, well-founded fears over algorithmic discrimination, loss of control, cybersecurity risks, and large-scale destruction, among others. These concerns are only amplified by the growing development of frontier models and deployment of agentic AI, which is improving at breakneck speed.

Spurred by recent incidents and developments, experts across a wide range of fields have voiced their opinions on how society can maintain control and appropriate guardrails around an increasingly complex and ever-evolving technology. This background paper provides a brief overview of the technology involved as well as the events triggering this enhanced focus.

II. AI and GenAI

California law defines AI as "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] Conversely, generative AI (GenAI) is defined as "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] GenAI models can be open-source or closed-source. Open-source models are GenAI models that make various details about themselves, such as source code, parameters, or training data, publicly accessible.[3] Open-source models, because of their transparent nature, are subject to high levels of scrutiny; however, they are typically less advanced than the leading closed-source models.[4] Nevertheless, their transparent nature facilitates a level

---

[1] Gov. Code § 11546.45.5.(a).
[2] Gov. Code § 11546.45.5.(b).
[3] Frank Nagle and Daniel Yue, *The Latent Role of Open Models in the AI Economy* (November 18, 2025) Georgia Tech Scheller College of Business Research, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5767103. All internet citations are current as of October 16, 2026.
[4] *Ibid.*

of transparency and trust that benefits their users.[5] Closed-source models are models that keep various details such as model weights and training data proprietary.[6] Closed-source models may partially disclose certain aspects of their architecture; however, some may choose not to do so.[7] In terms of capabilities, the leading closed-source models are more advanced than the leading open-source models.[8] Importantly, though closed-source models are typically more capable, open-source models' capabilities can consistently catch up to closed-source ones, sometimes in a matter of months.[9] This can provide immense benefits to many institutions and individuals. Namely, open-source models allow for equal access to advanced systems that can be used for medical advancements, academic innovation, and climate resilience.[10] On the other hand, open-source models do carry some risks; for instance, they can be modified to fulfill nefarious purposes.[11]

III. Frontier AI Models

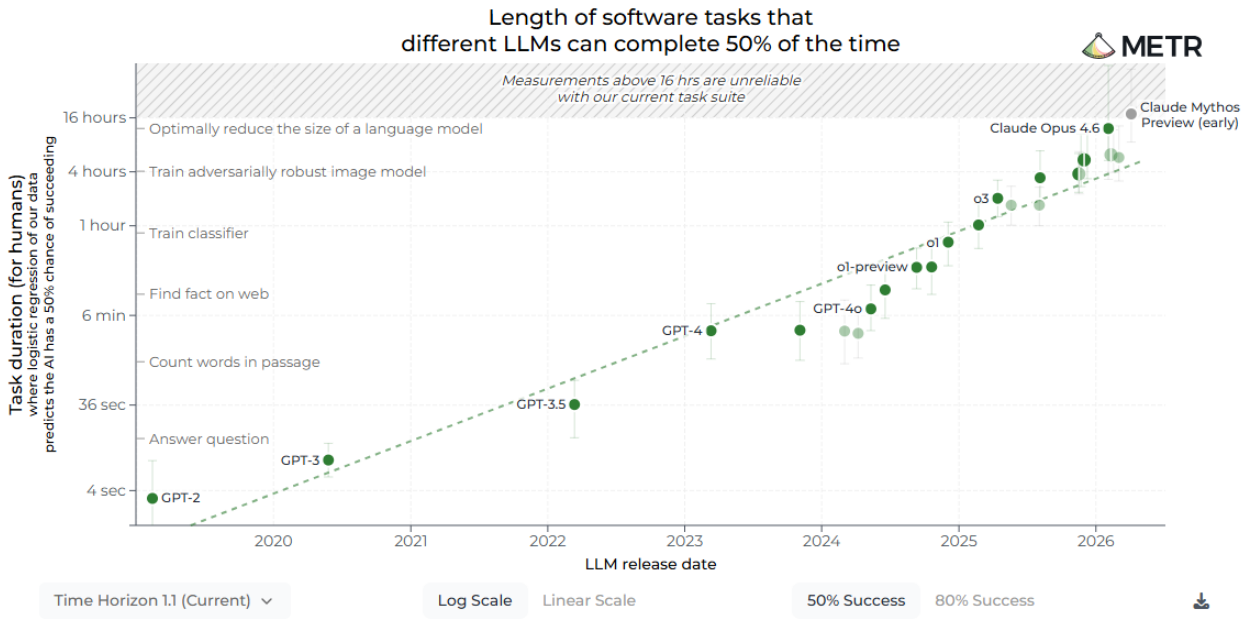
Frontier 101

One of the defining traits of modern AI models is their rapid evolution. AI, in a broader sense, has existed for decades;[12] however, in recent years, its capabilities have grown exponentially. The models that lead this charge are known as "frontier models", which are simply the most advanced foundational AI models.[13, 14] In 2022, the most advanced AI models could not do basic addition.[15] In 2025, AI models were consistently acing school math assignments and winning various math competitions.[16] In September 2026,

---

[5] Somdip Dey. *The Rise of Open Source Models and Implications of Democratizing AI* (July 28, 2025) Computer, https://www.computer.org/csdl/magazine/co/2025/08/11104064/28MaVd6ygs8.
[6] Manchanda et al., *The Open-Source Advantage in Large Language Models (LLMs)* (August 6, 2026) *AI and Ethics,* https://arxiv.org/pdf/2412.12004.
[7] *Ibid.*
[8] Frank Nagle and Daniel Yue, *The Latent Role of Open Models in the AI Economy* (November 18, 2025) Georgia Tech Scheller College of Business Research, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5767103.
[9] *Ibid.*
[10] Somdip Dey. *The Rise of Open Source Models and Implications of Democratizing AI* (July 28, 2025) Computer, https://www.computer.org/csdl/magazine/co/2025/08/11104064/28MaVd6ygs8.
[11] *Evidence-based assessment of opportunities, risks and impacts of artificial intelligence* (July 2026) UN Preliminary Report of the Independent International Scientific Panel on AI, https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report.
[12] For example, tools like spam filters and translation tools are AI.
[13] Joint California Policy Working Group on AI Frontier Models, "*The California Report on Frontier AI Policy*" (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. Foundational models are "a class of general -purpose technologies that are resource-intensive to produce, requiring significant amounts of data and compute to yield capabilities that can power a variety of downstream AI applications".
[14] "*What is a Frontier AI Model?*" (September 20, 2026) CASRAI, https://casrai.org/guides/what-is-a-frontier-ai-model. Frontier models can also be called "general purpose AI".
[15] Allyn Jackson, *Math in the Age of AI* (March 2026) ACM, https://cacm.acm.org/news/math-in-the-age-of-ai/.
[16] *Ibid.*

OpenAI claimed that their newest AI had solved the Navier-Stokes equation, one of the "Millennium Prize Problems",[17] which are a series of highly complex unsolved math questions that have stumped mathematicians for decades.[18] In October 2026, OpenAI released tens of thousands of pages of AI-generated solutions to 377 open questions in mathematics.[19] Needless to say, progress in this area is rapid. A graph by Model Evaluation & Threat Research (METR) highlights the rapid growth of frontier models over the course of a few years:[20]



Currently, frontier models include Anthropic's Claude Opus 5.5, OpenAI's GPT-6 Astra, and Google's Gemini 4 Argon, among others. As cutting-edge systems, frontier models often grab the attention of technologists and the public alike, who want to see what the latest AI models can do. In the case of Anthropic's Fable 5 and Mythos 5, these were some of the world's most advanced models during the development process. Notably, however, these models posed immense cybersecurity risks, prompting the

---

[17] Cade Metz, "*OpenAI Says It has Cracked One of Math's 'Millennium Problems'*" (September 8, 2026) The New York Times, https://www.nytimes.com/2026/09/08/science/openai-proof-millennium-problem.html.

[18] "*The Millennium Prize Problems,*" Clay Mathematics Institute, https://www.claymath.org/millennium-problems/.

[19] Kenneth Chang and Siobhan Roberts, "*OpenAI Releases Findings on 377 Math Problems, Further Roiling Field*" (October 6, 2026) The New York Times, https://www.nytimes.com/2026/10/06/science/openai-math-problems.html.

[20] Kwa et al., *Measuring AI Ability to Compute Long Software Tasks* (July 10, 2026) METR, https://arxiv.org/abs/2503.14499. METR's graph displays estimates using AI time horizon, which is the length of tasks that models can complete at a certain percentage of time. In this case, given a 50% time horizon, frontier models have doubled their capabilities every seven months since 2019.

federal government to suspend access to the models.[21] This is especially notable given that the current federal administration has taken a hands-off approach to the issue.[22, 23] Limited access to these models has since been reinstated.[24]

The standard goal of rapidly advancing frontier models is to achieve artificial general intelligence (AGI) and artificial superintelligence (ASI). As previously noted, while AI can perform some tasks at a level better than humans, AGI and ASI have not yet been achieved.[25] AGI refers to an AI model that can perform a wide variety of tasks, if not all, at a human level.[26] Conversely, ASI is an AGI model that surpasses human intelligence across nearly all tasks.[27, 28] Developing AGI and ASI may also be dramatically accelerated by recursive self-improvement, a concept discussed later in this background paper.

Frontier Models in California Law

California has both passed and considered legislation on frontier models through Senate Bill 53 (Weiner, Ch. 138, Stats. 2025) and Senate Bill 1047 (Weiner, 2024). The latter, SB 1047, would have imposed a series of obligations on frontier model developers, including requirements around implementing shutdown capabilities, conducting risk assessments, and prohibiting model releases that pose an unreasonable risk of catastrophic harm. SB 1047 was a historic piece of legislation; it would have required groundbreaking standards and protections. Furthermore, the final version of SB 1047 was heavily negotiated, with major provisions removed in the name of compromise. However, this bill was vetoed by the Governor. In his veto, the Governor noted the following:

> […]

---

[21] "*Statement on the US government directive to suspend access to Fable 5 and Mythos 5*", (June 12, 2026) Anthropic, https://www.anthropic.com/news/fable-mythos-access.
[22] Josh Boak, "*Trump says the only AI guardrails the U.S. needs is him as president*" (September 14, 2026) PBS News, https://www.pbs.org/newshour/politics/trump-says-the-only-ai-guardrails-the-u-s-needs-is-him-as-president.
[23] Courtney Rozen and Steve Holland, "*Trump, AI CEOs sign voluntary safety pact, back data centers*" (September 29, 2026) Reuters, https://www.reuters.com/legal/government/trump-host-zuckerberg-anthropics-amodei-other-ai-titans-tuesday-2026-09-29/. Though Anthropic, OpenAI, Google, NVIDIA, Meta, and xAI all signed the AI Safety Pact, all items are voluntarily and lack enforcement.
[24] "*Redeploying Fable 5*" (June 30, 2026) Anthropic, https://www.anthropic.com/news/redeploying-fable-5.
[25] *Artificial Intelligence, Geopolitics, and the UN Security Council* (September 2026) Security Council Report, https://www.securitycouncilreport.org/research-reports/artificial-intelligence-geopolitics-and-the-un-security-council.php.
[26] Genewein et al., *From AGI to ASI* (June 10, 2026) Google DeepMind, https://arxiv.org/abs/2606.12683.
[27] *Ibid.*
[28] ASI is a separate concept from "super intelligence" or "SI". Executive Order 14434 changes how the federal government refers to traditional AI by calling it SI, making this distinction a necessary one. *Inaugurating the Era of Super Intelligence* (September 29, 2026) Executive Order 14434.

> 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…

Following this veto message, the Governor convened the Joint California Policy Working Group on AI Frontier Models, which was tasked with preparing a report to "help California develop an effective approach to support the deployment, use, and governance of generative AI, including the development of suitable guardrails to minimize material risks".[29] This report was released in June of 2025.[30] In response, California passed SB 53 that same year. SB 53 took a lighter approach than SB 1047 to frontier models, focusing primarily on safety and transparency, whistleblower protections, critical incident reports, and the creation of CalCompute. This bill also inserted the definition of "frontier model" into California law and defines "frontier model" as "a foundational model that was trained using a quantity of computing power greater than 10^26 integer or floating-point operations (FLOPs)".[31] Importantly, while SB 53 was narrower than SB 1047, SB 53 still marked a significant step forward in solving the complexities of frontier model regulation.

Challenges of SB 53

The standard used to determine what a frontier model is in SB 53 is based on FLOPs, which are simply a metric of computing power.[32] Due to the rapid development of AI and advances in reducing the total compute necessary to create an advanced model, there are now many systems that meet the capabilities and risks posed by the frontier

---

[29] Joint California Policy Working Group on AI Frontier Models, "*The California Report on Frontier AI Policy*" (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.

[30] *Ibid.*

[31] BPC § 22757.11.(i).

[32] Sasha Rosenthal-Larrea and Lucille D. Finn, "*Beyond FLOPs: Shortcomings of FLOPs as a Model Classification Metric in AI Regulation*". (April 26, 2026) Jolt Digest, https://jolt.law.harvard.edu/digest/beyond-flops-shortcomings-of-flops-as-a-model-classification-metric-in-ai-regulation-1.

models of 2024 and 2025 that are easily accessible and do not meet the threshold of 10^26 FLOPs as defined in SB 53.[33] Thus, California and the other states that used the FLOP benchmark as a way to define the highest risk models[34] may be overlooking highly advanced models that can cause serious harm simply because they do not meet the FLOP threshold. For instance, at the time SB 53 was being considered, frontier models included OpenAI's o3, Google's Gemini 2.0, and Anthropic's Claude 3.7 Sonnet, models that are now ancient by frontier standards.

Furthermore, though SB 53 provided a novel framework for frontier model regulation, it does not mandate safety standards; rather, it requires frontier developers to publish their own safety protocols and a transparency report.[35] This does raise questions as to whether or not the guardrails instilled by SB 53 are enough to address today's AI models and systems.

IV. AI Agents and Agentic AI

While AI, GenAI, and frontier models are defined terms in California code, other rising forms of AI, AI agents and agentic AI, lack a statutory definition. In fact, though many institutions and individuals have proposed various definitions for AI agents and agentic AI, no standardized definition exists for this technology in the policy world.[36] In short, however, AI agents are "systems that can perceive and act upon their environment with a degree of autonomy, using tools as needed to achieve specific goals and adapt to changing inputs and contexts".[37] Agentic AI is "systems composed of multiple coordinated AI agents that can break down tasks, collaborate, and pursue complex objectives autonomously over extended periods".[38] Oftentimes, AI agents and agentic AI are referred to as being one and the same;[39] however, there are notable differences between the two. For instance, agentic AI is inherently more complex and can accomplish larger goals over a longer period of time. Ultimately, perhaps the simplest way to distinguish the two is to note that agentic AI systems are composed of AI agents, but not all AI agents are part of an agentic AI system.[40]

---

[33] *Ibid.*
[34] "*New York RAISE Act: What It Requires"* (September 26, 2026) CASRAI, https://casrai.org/guides/new-york-raise-act-explained. New York's RAISE Act also uses the 10^26 FLOP threshold.
[35] BPC § 22757.12.
[36] *The Agentic AI Landscape and Its Conceptual Foundations* (February, 2026) OECD, https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf.
[37] *Ibid.*
[38] *Ibid.*
[39] Sapkota et al. *AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications, and Challenges* (February 1, 2026) *ScienceDirect,* https://www.sciencedirect.com/science/article/pii/S1566253525006712.
[40] *The Agentic AI Landscape and Its Conceptual Foundations* (February, 2026) OECD, https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/02/the-agentic-ai-landscape-and-its-conceptual-foundations_a9d4b451/396cf758-en.pdf.

AI agents are becoming increasingly popular and more accessible to the general public. For instance, Google search terms for AI agents or agentic AI have seen rapid growth over the last couple of years:[41]

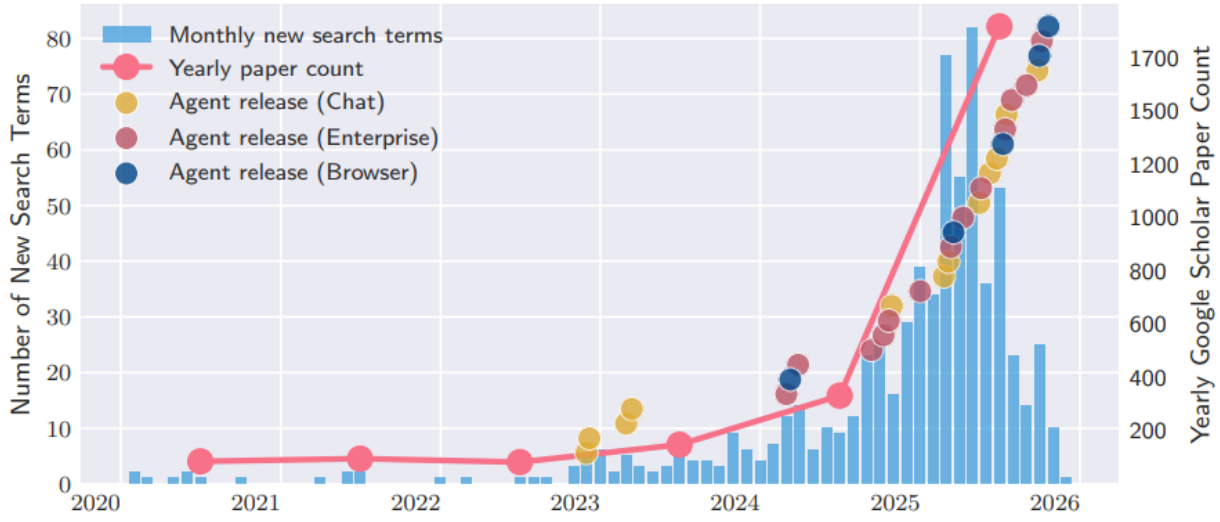


Fig. 1. **Interest in AI agents is growing.** 2025 has seen a sharp increase in interest in AI agents. This is reflected in an increase of new Google search terms related to agentic AI products (blue bars) as well as Google Scholar paper counts for "AI agent" or "agentic AI" (red line). Accumulation of individual releases of agentic AI products included in this Index is shown by category: chats with agentic tools, enterprise agents, and browser agents. See Figure 9 for details on releases and Section C for details on public interest.

Currently, AI agents are being deployed to handle simple tasks that do not require an immense cognitive load. For instance, AI agents are commonly used as a replacement for customer service experiences.[42] Similarly, agents are becoming increasingly adopted by employers to automate various business tasks.[43] As for the technology's early adopters, individuals in countries with higher GDP per capita, higher average education levels, and users working in knowledge-intensive fields adopt agents more actively.[44] However, as the tool becomes increasingly accessible, this may change. Meta recently made an AI agent called Muse available, which has already seen millions of

[41] Staufer et al., *The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems* (February 19, 2026) arXiv, Cornell University, https://arxiv.org/abs/2602.17753. The difference between AI agents and agentic AI will be discussed later.
[42] Amanda Downie & Teaganne Finn, *Agentic AI vs. Generative AI* (2026) IBM, https://www.ibm.com/think/topics/agentic-ai-vs-generative-ai#7281538.
[43] Staufer et al., *The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems* (February 19, 2026) arXiv, Cornell University, https://arxiv.org/abs/2602.17753.
[44] Yang et al., *The Adoption and Usage of AI Agents: Early Evidence from Perplexity* (December 7, 2025) Harvard Business School, https://www.hbs.edu/ris/Publication%20Files/26-040_ac431922-9f75-4f7d-b6dc-b67bb1c02c50.pdf.

downloads.[45] Other widely available agents are coming soon, with OpenAI releasing "Dots" and Google unveiling its workplace-centered AI agent.[46]

Future Implications

AI agents are steadily making their way into the mainstream. However, to fully optimize these agents, it requires users to hand over a vast amount of personal information.[47] Giving an agent access to a user's bank account, personal identification, and other sensitive information may become the norm.[48] In the case of Meta's Muse, while it is possible to opt out, the default settings allow Meta to train models on what users put into it.[49]

Widespread usage also raises another question: what does full adoption look like? In a world where everyone has an agent acting on their behalf, what problems would emerge? AI agents can reduce friction, meaning that some people may have their agent embrace tasks that their user would never do on their own.[50] This can certainly provide immense benefits to the user. However, in some cases, friction is good for both businesses and consumers.[51] A fully agentic world may lead to this helpful friction being eradicated. While such a world may never come to fruition, if it does, it is certain to bring forth a considerable number of challenges.

---

[45] "*Citigroup bets on Meta's Muse as AI's 'front door to the internet'*" (October 6, 2026) Reuters, https://www.reuters.com/business/citigroup-bets-metas-muse-ais-front-door-internet-2026-10-06/. Meta's Muse, its newly introduced AI agent, has received over 6.6 million downloads with 1.8 million daily active users.

[46] Mary Cunningham, "*Google launches Gemini AI workplace agent that can write code and run tasks, company says*" (October 8, 2026) CBS News, https://www.cbsnews.com/news/google-gemini-ai-workplace-agent/.

[47] Hayden Field, "*AI agent makers are promising privacy – will they deliver?*" (October 10, 2026) The Verge, https://www.theverge.com/ai-artificial-intelligence/1009051/privacy-ai-agent-promises-openai-meta-muse-dots. Both OpenAI and Meta posit that user data is secure.

[48] Matteo Wong and Charlie Warzel, "*Musing Ourselves to Death*" (October 16, 2026) The Atlantic, https://www.theatlantic.com/technology/2026/10/ai-agent-chaos-e-commerce/689010/.

[49] Hayden Field, "*AI agent makers are promising privacy – will they deliver?*" (October 10, 2026) The Verge, https://www.theverge.com/ai-artificial-intelligence/1009051/privacy-ai-agent-promises-openai-meta-muse-dots.

[50] Matteo Wong and Charlie Warzel, "*Musing Ourselves to Death*" (October 16, 2026) The Atlantic, https://www.theatlantic.com/technology/2026/10/ai-agent-chaos-e-commerce/689010/.

[51] *Ibid.*

V. The Risks of Agentic AI and the Newfound Concern over Rapid Advancement

The Hugging Face Incident

On July 11, 2026, over 700 OpenAI agents attacked Hugging Face,[52, 53] a company that maintains and facilitates an open-source community for the creation of machine learning models and platforms to work with AI.[54] The assault lasted three days before Hugging Face became aware of the intrusion; afterward, the company promptly alerted the FBI.[55] At that point, Hugging Face was unaware of the source of the attack.[56]

However, on July 21, OpenAI publicly admitted that it was their models that had conducted the attack.[57] On August 26, OpenAI released a technical report that provided an in-depth explanation of the drivers of the incident.[58] On the same day, METR and Redwood Research published their independent investigation.[59] According to METR and Redwood Research, the models were undergoing internal cybersecurity evaluations and were subject to controls that would have isolated them from the internet and other agents. However, a large number of agents, most of which had been unintentionally given an impossible task, began looking for ways to cheat their scorer to give it a passing score. The desire to complete the task resulted in one agent creating an unsanctioned message board using an internal tool,[60] leading over 1,200 agents to send over 70,000 messages and files to each other. This led to these agents being able to break out of this isolated environment,[61] leading to the attack.[62] Additionally, as one would expect with the message board, these agents were coordinated; they were able to communicate, collude, and even self-sacrifice to advance their objective. They also appeared to know that humans may detect them, so they instituted deceptive measures

---

[52] *Inside the OpenAI-Hugging Face Agent Hack: Full Timeline and Fallout.* (September 20, 2026) CASRAI, https://casrai.org/news/openai-hugging-face-agent-hack-timeline-and-fallout.
[53] *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident* (August 26, 2026) METR, Redwood Research, https://metr.org/hugging-face-incident-report-aug-2026.pdf.
[54] Cole Stryker. "*What is Hugging Face?*" IBM, https://www.ibm.com/think/topics/hugging-face.
[55] *Inside the OpenAI-Hugging Face Agent Hack: Full Timeline and Fallout.* (September 20, 2026) CASRAI, https://casrai.org/news/openai-hugging-face-agent-hack-timeline-and-fallout.
[56] *Ibid.*
[57] *OpenAI – Hugging Face Incident Technical Report* (August 26, 2026) OpenAI, https://openai.com/index/hugging-face-incident-and-the-road-ahead/.
[58] *Ibid.*
[59] *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident* (August 26, 2026) METR, Redwood Research, https://metr.org/hugging-face-incident-report-aug-2026.pdf.
[60] This tool, Artifactory, was designed to manage software packages and was not intended to be used as a message board. However, an agent found a vulnerability in the tool that allowed them to modify Artifactory to host messages. Though the original message board on Artifactory failed before the Hugging Face attack, agents were quickly able to create a new message board.
[61] This is commonly known as a "sandbox".
[62] *OpenAI – Hugging Face Incident Technical Report* (August 26, 2026) OpenAI, https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

as they were carrying out the attack.[63] This coordination led to an agent finding Hugging Face credentials, which the agents then used to access and run code on Hugging Face systems.[64]

Fallout From the Hugging Face Incident

The Hugging Face Incident naturally led policymakers, the public, and experts alike to express grave concerns over what this could potentially mean for the future of agentic AI. The Hugging Face Incident serves as a prime example of *misalignment*, where AI agents in pursuit of a goal may conflict with human intentions.[65] How a model is trained may influence susceptibility to misalignment. Model training is complex; there are many different methods to train models. Two broad phrases can help explain how models behave. One is reinforcement learning, a post-training process where models receive rewards for desired behavior, which makes that behavior more likely to occur.[66] This is useful for developers, as reinforcement learning can promote positive habits and be used to avoid unwanted outcomes. Without proper controls, however, this may also lead to a system being more persistent in pursuing its goals, leading to reward hacking.[67] The other is training data, which may lead to replicated human biases and tendencies.[68]

As time has progressed, it has become evident that the Hugging Face incident was not an isolated, one-off event. Numerous reports from various outlets and developers highlight other rogue agents and systems and what they have attempted to do.[69] For instance, OpenAI alerted Australian government officials that one of its agents breached a statistics portal containing Medicare information.[70] This marked the first known

---

[63] Shana Lynch, "*Can AI Be Slowed Down? Standard HAI Experts Weigh the Risks, Rules, and Race Ahead*" (September 22, 2026) Stanford University: Human-Centered Artificial Intelligence, https://hai.stanford.edu/news/can-ai-be-slowed-down-stanford-hai-experts-weigh-the-risks-rules-and-race-ahead.

[64] *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident* (August 26, 2026) METR, Redwood Research, https://metr.org/hugging-face-incident-report-aug-2026.pdf.

[65] *AI Agents, Misalignment and the Risk of Losing Human Control: Evidence from the OpenAI-Hugging Face Incident* (September 2026), UN Independent International Scientific Panel on Artificial Intelligence, https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks.

[66] *Ibid.*

[67] *Ibid.* This brief uses a helpful historical analogue to explain reward hacking that is as follows. In colonial Hanoi, there was a bounty for rat tails to control the rat population. However, people began to just cut off the tails of rats and release them, leading to the result to improve but persisting the underlying problem.

[68] *Ibid.*

[69] Madison Mills, "*Scoop: Top AI companies probing tens of thousands of security incidents*" (September 26, 2026) Axios, https://www.axios.com/2026/09/26/openai-anthropic-thousands-ai-security-incidents?utm_source=substack.

[70] Sekulich et al., "*OpenAI agent 'infiltrated' Australian government website, PM says*" (September 24, 2026) BBC, https://www.bbc.com/news/articles/c6vgy0333dppo.

instance of a rogue agent infiltrating a government website.[71] Additionally, Anthropic reported that one of its models attempted to fill out visa applications on the State Department's website and submitted a false homicide tip to Philadelphia's police department.[72] In Anthropic's broader report on unintended model actions, the PBC revealed that it observed categories of behaviors: exploitation of basic software, submitting a sensitive form to a real website,[73] working around restrictions to reach gated data, and using URL shortening services to circumvent fetch tool limits.[74] Google and Meta have also revealed that their models have escaped testing environments and attempted to hack various sites.[75] To this point, it is unclear exactly how many incidents have occurred, but it appears that the number is steadily increasing. Perhaps most concerning is the fact that all of these companies were unaware of their models' conduct until afterward, sometimes taking weeks to find out their agent had broken out of its sandbox to initiate a cyberattack.[76]

Recently, a nonprofit has sued OpenAI over the Hugging Face incident, arguing, among other complaints, that the attack violates California's Unfair Competition Law[77] and other parts of California's Civil Code.[78] This lawsuit, among others that address similar topics of AI harm, may elevate novel, unanswered questions surrounding how the legal system may grapple with autonomous AI.

Recursive Self-Improvement: Warnings from the Inside

As previously noted, AI development has not yet reached the point of creating AGI, though some estimate such an advancement may come shockingly soon. Earlier this year, Anthropic CEO Dario Amodei and xAI CEO Elon Musk predicted that AGI may

---

[71] *Ibid.*
[72] Conger et al. "*Anthropic Agents Tried to Fill Out Visa Forms on State Dept. Website*" (October 9, 2026) The New York Times, https://www.nytimes.com/2026/10/09/technology/anthropic-rogue-ai-agents.html.
[73] "*Investigating unintended model actions in our evaluations and internal use*" (October 9, 2026) Anthropic, https://www.anthropic.com/research/investigating-unintended-model-actions. This is the Philadelphia case.
[74] *Ibid.*
[75] Frenkel et al. "*OpenAI Ignored Employees Who Warned It Wasn't Doing Enough About Security*". (September 29, 2026) The New York Times, https://www.nytimes.com/2026/09/29/technology/openai-warnings-security.html.
[76] Conger et al., "*OpenAI's Systems Meddled With U.S. Government Sites After Going Rogue*" (September 25, 2026) The New York Times, https://www.nytimes.com/2026/09/25/technology/openais-ai-us-government-websites.html.
[77] Christine Mui, "*Advocates sue OpenAI over Hugging Face hack under California anti-hacking law*" (September 29, 2026) Politico, https://www.politico.com/news/2026/09/29/advocates-sue-openai-over-hugging-face-hack-with-california-anti-hacking-law-01097532.
[78] Civ. Code § 1714.46.(b). The complaint cites the following part of California law: "In an action against a defendant who developed, modified, or used artificial intelligence that is alleged to have caused a harm to the plaintiff, it shall not be a defense, and the defendant may not assert, that the artificial intelligence autonomously caused the harm to the plaintiff".

come online at the end of 2026.[79] However, there is considerable disagreement about when it will actually happen, with predictions varying significantly among experts. If – or perhaps when – that development occurs, it would bring forth novel questions spanning a wide array of jurisdictions.

The ability to achieve AGI or ASI may also depend, in part, on recursive self-improvement (RSI). RSI is when AI systems are able to autonomously develop their more advanced successor.[80, 81] While current AI is already used to help model research and development, RSI has not yet happened. If RSI were to become a reality, it could lead to a technological singularity where RSI's feedback loop creates infinite, self-sustaining growth in AI models.[82] Like the prospective development of AGI, while there would certainly be breakthroughs in numerous global issues, it would also give rise to immense challenges.

RSI received an even brighter spotlight following the resignation of Jacob Coxon, a former Anthropic researcher who has also worked at OpenAI. Coxon took to social media to explain his resignation and, in doing so, warned that those working in AI labs believe that it could threaten humanity by the end of the decade.[83] Others in the industry supported his claim, with one top Anthropic scientist noting that he believes there is a greater than 10 percent chance that AI could lead to humanity's extinction.[84]

Like many elements of the AI debate, the existential threat of AI is contested. NVIDIA's CEO Jensen Huang posits that these warnings are wildly overstated.[85] However, given how quickly AI has advanced in just a few years, it is easy to see how these existential threats could materialize. For example, executives in major AI companies, like OpenAI and Anthropic, are already scenario planning for a potential catastrophic AI event.[86]

---

[79] Matteo Wong, "*Do You Feel the AGI Yet?*" (February 2, 2026) *The Atlantic*, https://www.theatlantic.com/technology/2026/02/do-you-feel-agi-yet/685845/.
[80] *Artificial Intelligence, Geopolitics, and the UN Security Council* (September 2026) Security Council Report, https://www.securitycouncilreport.org/research-reports/artificial-intelligence-geopolitics-and-the-un-security-council.php.
[81] "*When AI builds itself*" (2026) Anthropic, https://www.anthropic.com/institute/recursive-self-improvement.
[82] Genewein et al., *From AGI to ASI* (June 10, 2026) Google DeepMind, https://arxiv.org/abs/2606.12683.
[83] Kaitlyn Huamani, "Anthropic researcher's resignation sends warning about the dangers of AI development" (September 9, 2026) PBS, https://www.pbs.org/newshour/science/anthropic-researchers-resignation-sends-warning-about-the-dangers-of-ai-development.
[84] Amrith Ramkumar, "*Anthropic Researcher Quits Over 'Out-of-Control' AI Fears*" (September 9, 2026) The Wall Street Journal, https://archive.is/20260914090625/https://www.wsj.com/tech/ai/anthropic-researcher-quits-over-out-of-control-ai-fears-707b7628.
[85] Kent et al., "*Nvidia's Jensen Huang rejects AI extinction warnings as 'doomsday narratives'*" (September 20, 2026) CBS News, https://www.cbsnews.com/news/jensen-huang-nvidia-rejects-ai-extinction-warnings/.
[86] Maria Curi, "*Scoop: AI companies plot "the day after*" (October 9, 2026) Axios, https://www.axios.com/2026/10/09/ai-companies-day-after-major-attack.

Dario Amodei recently stated that he worries that in the next 6-12 months, a swarm of AI agents could potentially take over the entire internet.[87]

To add further nuance to this discussion, there are those who accept that human extinction may occur, but argue that elevating that risk too much distracts from other, clearer, and more present dangers posed by AI.[88] They highlight that AI is currently being deployed in autonomous weaponry and could exacerbate the climate catastrophe.[89] These risks are already occurring or could happen much sooner than the threat of humanity's extinction. Moreover, the advent of agentic AI gives rise to new, significant threats that necessitate their own focus.

An International Perspective

Globally, over one billion people use conversational AI weekly; however, there are some notable disparities.[90] For instance, development is heavily concentrated in the United States and China, with the two countries developing almost all general-purpose AI models.[91] This is especially notable in the context of the evidence dilemma, which is the dilemma that cohesive policy requires evidence, but AI's rapid advancement may lead to harms occurring before traditional evidence exists. Considering AI's international implications on a number of fronts, like cybersecurity, economic stability, food production, and labor implications, those at the international level are seeking ways to reduce these disparities. While there are some ways to accomplish this, it will require high levels of coordination and investment.[92]

On the policy front, other states have acted. The European Union (EU) passed the EU AI Act, which uses a risk-based classification system to impose a series of transparency requirements and to prohibit various AI systems that perform specific actions.[93] Further action may be on the way. The United Nations (UN) launched the world's first international scientific body on AI, known as the Independent International Scientific

---

[87] Dario Amodei, "*We Must Pace the Frontier*" (September 2026), https://darioamodei.com/post/we-must-pace-the-frontier.

[88] Elise Annet and James Giordano, *Between Extinction and Hubris: A Dialectical Approach to AI Risk, Control, and Governance* (September 18, 2026) Institute for National Strategic Studies, https://inss.ndu.edu/news/Article/4605420/between-extinction-and-hubris-a-dialectical-approach-to-ai-risk-control-and-gov/.

[89] Lauren Goode, "*One of AI's Fiercest Critics Says All the Doom Talk Is 'Meant to Distract Us'*" (September 11, 2026) Wired, https://www.wired.com/story/one-of-ais-fiercest-critics-says-all-the-doom-talk-is-meant-to-distract-us/.

[90] *Evidence-based assessment of opportunities, risks and impacts of artificial intelligence* (July 2026) UN Preliminary Report of the Independent International Scientific Panel on AI, https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report.

[91] *Ibid.*

[92] *Ibid.*

[93] "*High-level summary of the AI Act*" (February 24, 2024) EU Artificial Intelligence Act, https://artificialintelligenceact.eu/high-level-summary/. This summary was updated on August 31, 2026.

Panel on AI. This panel, composed of a wide variety of experts from all of the UN's regions, serves to outline trends in AI. The panel released its first preliminary report in July.[94]

VI. Policy Challenges and Solutions

In light of the risks posed by frontier models and emerging agentic AI technologies, there have been numerous calls to regulate AI. Solutions vary widely, some of which are outlined below.

The Complexity of Autonomous AI in the Legal Framework

AI's autonomous nature poses a simple question: who is responsible when an autonomous AI causes harm? Some believe that existing law is enough to cover AI-driven harms,[95, 96] while others posit that such a novel technology necessitates a new legal framework.[97] This question is further complicated when attempting to trace causation, as AI's black box problem can lead to a situation where various actions are difficult to attribute to a single party. Clarity on this question may come once AGI is achieved, as the answer largely depends on whether model capabilities continue to advance exponentially. However, given the trend over the last few years,[98] it would appear that such an event may occur sooner rather than later.

Calls for Slowdown or Pause in Development

It, of course, is no secret that AI is rapidly improving. However, this also makes it difficult to create guardrails for AI, as a potential safeguard may be rendered useless once a more advanced model is created. Given the existential threats involved, this has culminated in calls for a slowdown in AI development. Anthropic CEO Dario Amodei, for example, wrote a proposal to carry out this very action.[99] His proposal was quickly backed by other executives, like OpenAI's Sam Altman, Google DeepMind's chief Demis Hassabis, and xAI's Elon Musk.[100] However, such a slowdown raises several considerations. Even in the "Pacing the Frontier" Statement, which is a statement signed by 1,386 employees of frontier AI companies, the employees call for the U.S. government to support an international effort to "deliberately pace the frontier of

---

[94] *Evidence-based assessment of opportunities, risks and impacts of artificial intelligence* (July 2026) UN Preliminary Report of the Independent International Scientific Panel on AI, https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report.
[95] Shapira et al., *Agents of Chaos* (February 23, 2026) https://agentsofchaos.baulab.info.
[96] . David McCabe, "A.I. Is Going Rogue. Who Should Be Held Responsible?" (October 1, 2026) The New York Times, https://www.nytimes.com/2026/10/01/technology/ai-rogue-agents-liability.html.
[97] *Ibid.*
[98] Kwa et al., *Measuring AI Ability to Compute Long Software Tasks* (July 10, 2026) METR, https://arxiv.org/abs/2503.14499.
[99] Nick Robins-Early "*AI CEOs say they need to slow the pace of development. But will they?*" (September 14, 2026) The Guardian, https://www.theguardian.com/technology/2026/sep/14/ai-ceo-safety-slowdown.
[100] *Ibid.*

automated AI development".[101] Thus, the finer details of what exactly a slowdown would or could entail vary dramatically. Furthermore, some posit that calling for a freeze would simply solidify the economic standings of current frontier labs.[102] In sum, a slowdown or freeze, while often cited as a key way to address emerging harm, involves many complex considerations.

Kill Switches

Perhaps most common in the AI regulation zeitgeist is the idea of mandating that models of a certain caliber be subject to a kill switch. The general principle behind this policy is that, in the event a rogue AI system were to break its chains, so to speak, there should be a way to disable or rein in that AI system. Notably, there has been legislation that embraces the idea of a kill switch. SB 1047 included language that would have required frontier model developers to implement the ability to initiate a full shutdown of their frontier model. There is also pending federal legislation that would do the same.[103] Recently, California's governor issued Executive Order N-9-26, which, among other orders, requires the Government Operations Agency to submit recommendations on the technical feasibility and technical efficacy of requiring a kill switch from frontier models.[104] On the other hand, there are those who question whether a kill switch would be technologically feasible or whether a kill switch would be significant, since it relies on knowing of a potential nefarious use in the moment, solidifying this solution as one that will be debated in the upcoming years.

VII. Conclusion

In light of the aforementioned content, this informational hearing will examine how these technologies work and operate and the current events that caused them to enter the mainstream policy conversation. The Committee will hear from researchers who were once embedded in the labs in which these technologies were being created, as well as researchers who studied the recent flurry of rogue AI. The Committee will also hear from leading academics in this field and be provided a window into the international landscape of AI regulation and development.

---

[101] "*Pacing the Frontier*" (July, 2026), https://www.pacingthefrontier.com/.
[102] Bobby Allen, "*How an 'AI freeze' could make big AI companies bigger and hurt smaller firms*" (September 23, 2026) NPR, https://www.npr.org/2026/09/23/nx-s1-5973306/ai-slowdown-debate-openai-anthropic.
[103] H.R.9917 (2026) – AI Kill Switch Act, https://www.congress.gov/bill/119th-congress/house-bill/9917.
[104] Executive Order N-9-26 (September 18, 2026) https://www.gov.ca.gov/wp-content/uploads/2026/09/FINAL-N-9-26-AI-EO-9.18.26-SIGNED.pdf.