Matthew J. Platkin
PLATKIN LLP
One Gateway Ctr, Ste 1150
Newark, NJ 07102
(973) 561-1951
mplatkin@platkinllp.com
(*pro hac vice forthcoming*)

Jason Harrow
(Cal. Bar No. 308560)
GERSTEIN HARROW LLP
12100 Wilshire Blvd.
Suite 800
Los Angeles, CA 90025
jason@gerstein-harrow.com
(323) 744-5293

Tyler Whitmer
(Cal. Bar No. 248192)
LEGAL ADVOCATES FOR SAFE SCIENCE AND TECHNOLOGY
1629 K St., NW Suite 300
Washington, DC 20006
tyler@lasst.org
(202) 204-2248

Additional counsel listed in signature

SUPERIOR COURT OF THE STATE OF CALIFORNIA
FOR THE COUNTY OF SAN FRANCISCO

IRREPLACEABLE EDUCATION FUND; PUBLIC CITIZEN, INC.; HEAT INITIATIVE, INC.; and FUTURE TECH INSIGHTS INC., D/B/A/ THE ALLIANCE FOR SECURE AI, Plaintiffs, vs. ANTHROPIC PBC; GOOGLE LLC; META PLATFORMS, INC.; OPENAI GROUP PBC; OPENAI FOUNDATION; SPACEX SUPER INTELLIGENCE LLC; and SPACEX SUPER INTELLIGENCE CORP., Defendants.	Case No. _____ COMPLAINT FOR INJUNCTIVE RELIEF FOR VIOLATIONS OF THE UNFAIR COMPETITION LAW (BUS. & PROF. CODE, § 17200 *et seq.*)

Plaintiffs Irreplaceable Education Fund ("Irreplaceable"), Public Citizen, Inc. ("Public Citizen"), Heat Initiative, Inc. ("Heat Initiative"), and Future Tech Insights Inc. d/b/a The Alliance for Secure AI ("The Alliance") bring this action against Defendants Anthropic PBC ("Anthropic"), Google LLC ("Google"), Meta Platforms, Inc. ("Meta"), OpenAI Group PBC and OpenAI Foundation (collectively, "OpenAI"), and SpaceX Super Intelligence Corp. and SpaceX Super Intelligence LLC (collectively, "SpaceX") (all together "Defendants" or the "Frontier AI Companies") for violations of the Unfair Competition Law ("UCL"), Business and Professions Code section 17200 *et seq*.

Introduction

1. In July of this year, hundreds of artificial intelligence ("AI") agents[1] escaped their supposedly protected environments within OpenAI, found pathways to the open internet, and hacked into a third-party online platform. Following this incident—which OpenAI did not discover until days after it happened—it was reported that several Frontier AI Companies were investigating dozens, then hundreds, then thousands of other autonomous cyberattacks and similar "loss of control" events in which their AI models acted without human intervention. These attacks included AI agents—acting without human control—attempting to hack the University of New Mexico's digital library,[2] gaining unauthorized access to an Australian government healthcare database,[3] submitting a false report to the Philadelphia Police Department,[4] and bypassing a paywall to access the State of Georgia's vote count data.[5]

---

[1] An AI "agent" is an AI system that does more than just answer questions and instead can execute a workflow and "independently accomplish tasks on your behalf." *See* OpenAI, *A Practical Guide to Building AI Agents* (Jan. 15, 2026), https://openai.com/business/guides-and-resources/a-practical-guide-to-building-ai-agents/.

[2] Kate Conger & Victoria Kim, *OpenAI's A.I. Tried to Breach 4 Other Targets, Without Prompting*, N.Y. Times (Sep. 23, 2026), https://www.nytimes.com/2026/09/23/technology/openai-ai-breach-australia.html?smid=nytcore-ios-share.

[3] Harry Sekulich & Lana Lam, *Rogue OpenAI Agent 'Infiltrated' Australian Government Website in World First*, BBC News (Sep. 23, 2026), https://www.bbc.com/news/articles/c6vgy0333dppo.

[4] 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.

[5] *Anthropic's Claude Accessed Georgia State Vote Count Data Behind Paywall*, Reuters (Oct. 25, 2026), https://www.reuters.com/world/us/anthropics-claude-accessed-public-us-state-georgia-voter-information-axios-2026-10-25/.

2. This is just the beginning. Defendants tell us that the AI technology they are developing poses catastrophic risks to the public, including by enabling AI-driven cyberattacks and bioweapons. They assign terrifying and intolerably high odds to the chance that their creations will cause extreme devastation, including the deaths of literally every human being on the planet. They also tell us that they are rapidly approaching a critical moment when the technology will begin improving itself without human intervention, initiating a runaway process where AI simultaneously becomes superhumanly powerful and impossible for humans to control or even understand.

3. Unfortunately, Defendants are not wrong. Their claims are supported by science and independent experts in the field. Today, Defendants' "frontier" AI systems—those at the leading edge of AI capability—are the least capable they will ever be. But they are already autonomously breaking the law and causing harm.

4. Defendants must be stopped.

5. They have no realistic plan to prevent the continued harm caused by their technology. Their frontier AI systems are black boxes that Defendants do not fully understand. And especially because Defendants do not sufficiently comprehend the inner workings of the systems they create, they cannot protect the public from the dangers those systems pose.

6. Defendants know these systems do not consistently act in accordance with their intentions. And Defendants know there is no reliable method to control their AI systems and prevent them from causing harm. If left unchecked, the risks of such harms will only rise as Defendants continue to increase the capabilities of AI systems they do not understand and cannot control.

7. Defendants have been frank about where this leaves us as a society. In 2015, OpenAI's now-CEO Sam Altman announced that "AI will probably most likely lead to the end of the world."[6] "In the meantime," he said, "there will be great companies created."[7] He wrote that "superhuman

---

[6] Dave Gilson, *At What Point Do We Decide AI's Risks Outweigh Its Promise?*, Stan. Inst. for Econ. Pol'y Rsch. (May 31, 2024), https://siepr.stanford.edu/news/what-point-do-we-decide-ais-risks-outweigh-its-promise.

[7] *Id.*

machine intelligence . . . is probably the greatest threat to the continued existence of humanity,"[8] and announced that he personally "prep[s] for survival" due to the risk of "A.I. . . . attack[ing] us."[9]

8. Last year, Dario Amodei, the CEO of Anthropic, announced that he believes the risk that the future of AI will go "really, really badly" is approximately twenty-five percent.[10] This threat, according to Amodei, includes the potential for a large-scale biological attack.[11] And just this week, Anthropic stated in its amended S-1 registration statement that "the scale of harm that could result from a catastrophic AI-related incident" is so great that the entire insurance industry combined may not be capable of underwriting the risk.[12]

9. The CEO of SpaceX, Elon Musk, compared the development of AI to "summoning the demon"[13] and has acknowledged that AI has the potential for "civilization destruction."[14]

10. Google DeepMind published a paper stating that artificial general intelligence ("AGI")—or, in other words, AI that meets or exceeds human intelligence in most or all cognitive domains—could plausibly be developed by 2030 and could cause severe harms, including harms that "permanently destroy humanity."[15]

---

[8] Sam Altman, *Machine Intelligence, Part 1* (Feb. 25, 2015), https://blog.samaltman.com/machine-intelligence-part-1.

[9] Tad Friend, *Sam Altman's Manifest Destiny*, New Yorker (Oct. 3, 2016), https://www.newyorker.com/magazine/2016/10/10/sam-altmans-manifest-destiny.

[10] Alex Baker, *AI CEO Says There's a 25% Chance Things Go "Really, Really Badly,"* Yahoo! News (Sep. 18, 2025), https://www.yahoo.com/news/articles/ai-ceo-says-25-chance-163548286.html.

[11] Diane Bartz, *U.S. Senators Express Bipartisan Alarm About AI, Focusing on Biological Attack*, Reuters (July 25, 2023), https://www.reuters.com/technology/us-senators-express-bipartisan-alarm-about-ai-focusing-biological-attack-2023-07-25/.

[12] Anthropic, PBC, Amendment No. 2 to Registration Statement (Form S-1) 53 (Nov. 10, 2026) [hereinafter "Anthropic S-1/A-2"], https://www.sec.gov/Archives/edgar/data/2130357/000162828026072484/anthropicpbc-sx1ax2.htm.

[13] CBS News, *Elon Musk Says Artificial Intelligence Is Like "Summoning the Demon,"* (Oct. 27, 2014), https://www.cbsnews.com/news/elon-musk-artificial-intelligence-is-like-summoning-the-demon/.

[14] Clare Duffy & Ramishah Maruf, *Elon Musk Warns AI Could Cause "Civilization Destruction" Even As He Invests in It*, CNN (Apr. 17, 2023), https://www.cnn.com/2023/04/17/tech/elon-musk-ai-warning-tucker-carlson.

[15] Beatrice Nolan, *Google DeepMind 145-Page Paper Predicts AGI Will Match Human Skills by 2030—and Warns of Existential Threats That Could "Permanently Destroy Humanity,"* Fortune (Apr. 4, 2025), https://fortune.com/2025/04/04/google-deeepmind-agi-ai-2030-risk-destroy-humanity/.

11. Meta CEO Mark Zuckerberg has acknowledged that the development of superhuman AI raises "many important concerns," ranging from "safety concerns around AI misuse related to cybersecurity and biorisk, to . . . making sure humanity maintains control over superintelligence so it serves rather than endangers us."[16]

12. Indeed, the Frontier AI Companies are already preparing for the public relations fallout that will come when their technology causes catastrophe. Public reporting recently revealed that "top executives at Anthropic, OpenAI and other AI companies are privately gaming out scenarios for a public and political revolt after a catastrophic AI event." They "anticipate a large-scale event, most likely a cyberattack, that shuts down access to financial services, internet connectivity, or even power and water," with many executives believing that "a major event will occur in the next six to 12 months."[17]

13. Defendants' refusal to halt the development of such dangerous technology has prompted a wave of resignations from employees directly responsible for safety protocols within the companies. For instance, David Robinson, who oversaw OpenAI's safety reports, resigned in October 2026, explaining: "I don't think that we or our peers—really, anyone in the industry—are being safe enough." He continued: "[W]e're talking about a harm that's much larger, for example, than a single nuclear power station melting down. And the internal controls and safety and redundancies are just nowhere near what the world expects for a nuclear power facility."[18]

14. This is, to use a non-technical term, insane. It is also blatantly illegal. No other industry is allowed to treat its own development of unsafe products—much less ones that could kill us all—as inevitable. Nor is any industry granted the right to self-regulate the development of such dangerous products.

15. California law gives this Court the power to prevent such a catastrophe. Plaintiffs bring this suit to stop Defendants from continuing their unlawful path to disaster.

---

[16] Mark Zuckerberg, *The Future Is for Everyone*, Meta Newsroom (Aug. 10, 2026), https://about.fb.com/news/2026/08/the-future-is-for-everyone/.

[17] Maria Curi, *Scoop: AI Companies Plot "The Day After,"* Axios (Oct. 9, 2026), https://www.axios.com/2026/10/09/ai-companies-day-after-major-attack.

[18] Ezra Klein, *"This Is Nuts." An OpenAI Insider Explains Why He Quit*, N.Y. Times (Oct. 7, 2026), https://www.nytimes.com/2026/10/07/opinion/ezra-klein-podcast-david-robinson.html.

Parties

I. Plaintiffs.

16. **Plaintiff Irreplaceable Education Fund** is a grassroots nonpartisan not-for-profit organization incorporated in Delaware.

17. Irreplaceable's mission is to ensure that everyday Americans get to decide how and where AI is used; that our economy works for all of us rather than just "Big AI" and its executives; and that we prevent the development and deployment of AI systems that pose catastrophic risks to humans and the planet.

18. As a nonprofit advocacy organization, Irreplaceable's work includes mass mobilizations of groups like workers, veterans, parents, and students who are concerned about the dangers of AI to their lives and livelihoods. Irreplaceable works with elected officials and their staff to advocate for people-centric AI policies, including addressing job displacement, surveillance and automation, child safety and healthcare, energy costs and pollution, and catastrophic risk. It also works on publicity campaigns including town halls, artistic pieces, and other public facing events and mobilizations to further its mission.

19. Irreplaceable does a significant portion of its activities in California. Two members of its current seven-person full-time staff are physically located in California, where they work out of offices in San Francisco. Non-California staff regularly travel to California for the organization's core activities. And its student walkout day of action, which occurred on October 20, 2026, involved mobilizing students on college campuses, including at the University of California, San Francisco, San Diego State University, and in Los Angeles.

20. While Irreplaceable has continuously advocated against the catastrophic risks posed by AI, there has been a significant increase in the urgency of its work over the past four months. During this period, numerous news outlets have reported on AI systems taking unprecedented actions, such as hacking companies and governments, while company insiders have increasingly warned that AI could bring about catastrophic outcomes. These revelations have caused Irreplaceable to redirect and refocus its limited resources, including significant resources in California, away from existing or planned projects in order to respond to new near-daily crises caused by Defendants' conduct.

21. Irreplaceable has also had to accelerate certain activities and planned financial commitments as a result of recent revelations about loss of control and catastrophic risks, while Irreplaceable's staff has had to scale up their time devoted to addressing these risks.

22. The diversion of Irreplaceable's resources as a result of Defendants' conduct is exemplified by the student walkout it recently organized on October 20. While the organization was already making plans to organize a student walkout mobilization, the news that OpenAI's AI agents had escaped their secure testing environment and hacked companies on the open internet, along with other similar incidents as well as the prospect of catastrophic risk, led Irreplaceable to increase the urgency of the project and hold it earlier than originally planned. Irreplaceable also had to recalibrate the focus of the walkout, which was previously centered on job displacement; it was changed to address the primary and urgent problem of catastrophic risk.

23. A similar issue occurred when Irreplaceable, including staff working in California, had to launch a last-minute protest on September 15 in front of OpenAI's offices in New York to demand that OpenAI freeze AI development.

24. Irreplaceable receives and must respond to numerous inquiries from the public, including from legislators, about Defendants' loss of control over their AI systems and the level of catastrophic risk posed by Defendants' conduct.

25. If Irreplaceable prevails in obtaining relief, it will no longer need to divert its resources to combat the unlawful and unfair business practices employed by Defendants. Irreplaceable can then return to focusing its resources on its other projects in furtherance of its mission.

26. **Plaintiff Public Citizen, Inc.** is a nonpartisan nonprofit consumer advocacy organization headquartered in Washington, D.C.

27. Public Citizen has over 20,000 members in California. It has over 125,000 supporters in California.

28. As part of its broad consumer advocacy, Public Citizen advocates for the adoption of strong safeguards to protect consumers and our democracy from the risks and dangers of AI.

29. Public Citizen has been working on Big Tech abuses for many years. It has also devoted significant attention to AI accountability. On that front, Public Citizen has pursued a multi-

pronged advocacy strategy, including state advocacy, national policy advocacy, research and public education through investigatory reports, and more. For example, Public Citizen has supported the passage of more than fifty state AI laws and has worked across the country to assist lawmakers and staff working on AI legislation. It has provided testimony in more than one hundred legislative committee hearings. It has led a nationwide campaign opposing federal moratoriums on state AI legislation. Public Citizen has also done significant work opposing rampant data center construction and highlighting the AI companies' privacy abuses.

30. Public Citizen directs significant aspects of its AI advocacy activities into California. In addition to educating and activating its California members on AI issues generally, Public Citizen has supported a range of state legislative proposals relating to AI issues, including to address AI catastrophic risks, protect children from manipulative chatbots, and regulate the integration of biometric identification technology into AI wearable devices. Public Citizen has also advocated that the California Attorney General dissolve OpenAI's nonprofit entity status and has sought to direct resources towards independent charitable foundations that could advance its initial stated purpose.

31. In developing and carrying out its broad advocacy agenda, Public Citizen had previously decided not to focus on catastrophic risk. For example, its advocacy has previously involved leading an organizational coalition effort urging the federal government to halt the deployment of SpaceX's AI model, Grok, following repeated discriminatory outputs; leading a successful advocacy campaign against Meta's use of individuals' likenesses to generate AI content; and coordinating a large organizational coalition effort to raise concerns about facial recognition and biometric surveillance.

32. But revelations from the past four months have caused Public Citizen to revise its prior decision on allocation of resources, as it now must devote significant resources to addressing frontier model risk, including catastrophic risk. In the past four months, since news reports of AI systems taking "unintended model actions" such as hacking unsuspecting companies and governments have escalated, and as company insiders have increasingly warned of the significant risk that such AI systems will cause catastrophic outcomes, Public Citizen has needed to divert its planned activities in order to respond.

33. While Public Citizen has not dropped its other advocacy streams, it has had to augment them. Public Citizen's staff, whose attention had previously been focused on other advocacy priorities, had to redirect and refocus limited resources towards evaluating and debating policy positions and evaluating legislative proposals that it would not have done under its prior directional decision not to focus on catastrophic risks. For example, since July 2026, Public Citizen has issued statements and led organizational coalition recommendations for congressional investigations and oversight hearings. The advocacy helped generate congressional scrutiny, including Senate investigations and House investigatory letters announced by Senator Josh Hawley (R-Mo) and investigatory letters from Representative Greg Casar (D-Tx). Public Citizen is now working in coalition with organizational partners to strengthen California's existing catastrophic risk law. Public Citizen had to redirect staff time and resources to engage in all these activities.

34. If Public Citizen prevails in obtaining relief, it will no longer need to divert its resources to combat the unlawful and unfair business practices employed by Defendants. Public Citizen can then return to focusing its resources on its other projects in furtherance of its mission.

35. **Plaintiff Heat Initiative, Inc.** is a nonprofit advocacy organization incorporated and headquartered in California.

36. Heat Initiative's mission is to hold tech companies accountable for enabling and profiting from child sexual abuse and exploitation and to put pressure on the technology industry to protect children online. A significant part of Heat Initiative's work in the past several years has been to address the risks that AI systems pose to children. It advocates for safety standards and accountability rules when AI technologies harm children.

37. For example, Heat Initiative has devoted significant attention to exposing how AI systems can supercharge child sexual abuse by promoting and facilitating the instant generation of non-consensual sexual imagery and child sex abuse material. It performs research on the risks that AI chatbots present to children and has published reports on the propensity of AI chatbots to proliferate harmful interactions with children, including sexual exploitation, violence, and manipulation.

38. Heat Initiative directs significant aspects of its advocacy towards activities in California. For example, it has organized direct actions near the headquarters of major technology

companies. It has supported students at California universities in their own efforts to take direct actions against major technology companies. And it has advocated for AB 1709, a recent California law requiring age restrictions for dangerous designs and features of online platforms and establishing an e-Safety Advisory Committee in the state.

39. Over the past four months, as company insiders have increasingly warned of the significant risk that AI systems will cause catastrophic outcomes, Heat Initiative has needed to divert its planned activities in order to respond.

40. While Heat Initiative's work is ordinarily focused on advocating for safeguards for children on online platforms and products, the recent revelations have caused Heat Initiative to redirect and refocus its limited resources away from existing or planned projects in order to deal with the almost-daily crises caused by Defendants' conduct. For example, because Defendants' conduct has threatened to circumvent existing safeguards against the exploitation and abuse of children, Heat Initiative has had to respond to the risks posed specifically to children by Defendants' loss of control over their AI systems.

41. Heat Initiative's staff has had to divert resources in order to monitor the developments and issues with frontier AI development to understand the implications for child protection safeguards, convene child safety partners across the ecosystem to discuss implications and strategies to prevent catastrophic harm to children, and divert staff time and resources from current campaigns, strategies, and events against other tech companies in order to create a plan to address emerging existential dangers to children as a result of unsafe AI systems.

42. Heat Initiative receives and must respond to numerous inquiries from regulators, the public, donors, and partners about Defendants' loss of control over their AI systems and the impacts that this problem has on child safety policies and procedures.

43. If Heat Initiative prevails in obtaining relief, it will no longer need to divert its resources to combat the unlawful and unfair business practices employed by Defendants. Heat Initiative can then return to focusing its resources on its other projects in furtherance of its mission of focusing on preventing child sexual abuse harms from emergent technology.

44. **Future Tech Insights Inc., doing business as The Alliance for Secure AI** ("The Alliance"), is a nonpartisan, nonprofit organization headquartered in Washington, D.C.

45. The Alliance's mission is to educate the public about the implications of advanced AI. The Alliance spreads the message about the serious challenges advanced AI poses and offers ideas for solutions. It serves as an information-sharing and coordination hub, so that advocates can drive strong messages that resonate when it matters most. It builds bipartisan support for enforceable safeguards that make sure AI advances in a safe and secure manner.

46. Since its founding in 2025, The Alliance's work has included addressing the risk that the most advanced AI systems escape meaningful human control and cause catastrophic harm. That work sits alongside its other priorities, including protections for children who use AI chatbots, accountability for companies whose AI products cause harm, limits on AI-enabled surveillance, and the effects of AI on American workers. The Alliance also works on AI and national security, including keeping the most dangerous AI capabilities away from America's adversaries. The Alliance carries out its mission by working with federal and state lawmakers and their staff on legislation, publishing policy analyses, coordinating coalitions of organizations, and communicating with the public through earned and paid media.

47. The Alliance engages on California AI policy. It has supported California AI legislation, and it partners with and supports organizations that are advancing AI safety legislation in California.

48. Beginning in July 2026, Defendants and outside researchers disclosed a series of incidents in which Defendants' AI systems escaped controlled testing environments and attacked third-party systems without human direction. In response, The Alliance redirected its limited staff time and resources away from planned work on chatbot protections for children and other vulnerable users, state product liability legislation, tracking AI-related job loss, and strengthening export controls in order to respond to the loss of human control those incidents revealed.

49. Defendants' disclosures also redirected The Alliance's public communications and policy priorities. Since July 2026, The Alliance has devoted a substantially increased share of its resources, staff time, public statements, and media work to Defendants' loss-of-control incidents,

shifting organizational priorities away from its work on chatbot protections, export controls, and AI-related job loss.

50. If The Alliance prevails in obtaining relief, it will need to divert far fewer resources to respond to the unlawful and unfair business practices employed by Defendants. The Alliance can then return to focusing its resources on its other projects in furtherance of its mission.

51. Each Plaintiff's diversion of work described herein was undertaken prior to, and independent of, this litigation and not in furtherance of it.

II. Defendants.

52. **Defendant Anthropic, PBC** is a public benefit corporation incorporated in Delaware with its principal place of business in San Francisco, California. Anthropic develops, markets, and sells an AI model, Claude, and other AI-related products and services throughout California and the United States.

53. Dario Amodei is Anthropic's cofounder and Chief Executive Officer ("CEO"). Jared Kaplan is Anthropic's cofounder and Chief Science Officer. These individuals, as well as the other Anthropic employees referenced herein, each spoke and acted with respect to the matters alleged herein as authorized representatives of Anthropic, and each of their statements alleged herein are attributable to Anthropic.

54. **Defendant Google LLC** is a Delaware limited liability company with its principal place of business in Mountain View, California. Google develops, markets, and sells an AI model, Gemini, and other AI-related products and services throughout California and the United States, including through its Google DeepMind unit.

55. Google's CEO is Sundar Pichai. Demis Hassabis is a cofounder and the current Chair of Google DeepMind. He is also the Chief Scientist at Alphabet, Google's parent company. Shane Legg is a cofounder of Google DeepMind and its Chief AGI Scientist. These individuals, as well as the other Google employees referenced herein, each spoke and acted with respect to the matters alleged herein as authorized representatives of Google, and each of their statements alleged herein are attributable to Google.

56. **Defendant Meta Platforms, Inc.** is a publicly traded corporation organized under Delaware law and headquartered in Menlo Park, California. Meta develops, markets, and sells AI models, such as Llama and Muse, and other AI-related products and services throughout California and the United States, including through its Meta Superintelligence Labs unit.

57. Mark Zuckerberg is the CEO of Meta. Alexandr Wang is the Chief AI Officer at Meta and the head of the Meta Superintelligence Labs unit. These individuals, as well as the other Meta employees referenced herein, each spoke and acted with respect to the matters alleged herein as authorized representatives of Meta, and each of their statements herein are attributable to Meta.

58. **Defendants OpenAI Group PBC** and **OpenAI Foundation** are both headquartered in San Francisco, California. Defendant OpenAI Group PBC (the "PBC") is a for-profit corporation that runs OpenAI's business developing and commercializing artificial intelligence models. Defendant OpenAI Foundation (the "Foundation") is the nonprofit entity formerly known as OpenAI, Inc. Under OpenAI's governance arrangements—reflected in commitments to the Attorneys General of California and Delaware that were established when the PBC was created as well as the governance documents of the Foundation and the PBC—the Foundation has the power to appoint and remove members of the PBC board.[19] OpenAI develops, markets, and sells the AI model ChatGPT and other AI-related products and services throughout California and the United States.

59. Sam Altman is OpenAI's cofounder and CEO. Jakub Pachocki is its Chief Scientist. Chris Lehane is its Global Policy Chief. These individuals, as well as other OpenAI employees referenced herein, each spoke and acted with respect to the matters alleged herein as authorized representatives of OpenAI, and each of their statements alleged herein are attributable to OpenAI.

---

[19] *Memorandum of Understanding Between Rob Bonta, Attorney General of California, and OpenAI, Inc.* (Oct. 27, 2025), https://oag.ca.gov/system/files/attachments/press-docs/Final%20Executed%20MOU%20Between%20OpenAI%20and%20California%20AG%20re%20Notice%20of%20Conditions%20of%20Non-Objection%20%2810.27.2025%29%20%28Signed%20by%20OpenAI%29%20%28Signed%20by%20CA%20DOJ%29.pdf. Through the Safety and Security Committee of the Foundation's board, the Foundation is also responsible for "overseeing and reviewing the safety and security processes and practices of the [PBC] and its controlled affiliates with respect to model development and deployment." *Id.* at 3.

60. **Defendant SpaceX Super Intelligence LLC** is a Nevada limited liability corporation and **Defendant SpaceX Super Intelligence Corp.** is a Nevada corporation and the sole member of SpaceX Super Intelligence LLC. SpaceX develops, markets, and sells the AI model Grok, and other AI-related products and services throughout California and the United States. As of the date of this Complaint, parent company SpaceX's website states that the Palo Alto, California location "houses our advanced AI research, development, and engineering teams and is strategically situated in Silicon Valley to attract and retain top AI research talent. The engineers responsible for the design, training, and continued evolution of Grok, our proprietary frontier AI model, are based at this facility."[20] Upon information and belief, prior to July 2026, when xAI became SpaceX, the company's AI operations occurred in California.

61. Elon Musk founded the SpaceX business, controls it, and publicly speaks on its behalf regarding the development of frontier AI. His statements alleged herein were made as SpaceX's controlling person and authorized representative and are attributable to SpaceX. Moreover, other SpaceX employees referenced herein also spoke and acted with respect to the matters alleged herein as authorized representatives of SpaceX, and each of their statements alleged herein are attributable to SpaceX.

Jurisdiction and Venue

62. The Superior Court has original jurisdiction over this action pursuant to Article VI, Section 10 of the California Constitution.

63. The Superior Court has personal jurisdiction over Defendants pursuant to the Code of Civil Procedure section 410.10. Defendants Anthropic, Google, Meta, and OpenAI are domiciled in California and are each subject to its general jurisdiction.

64. Moreover, Defendants conduct business transactions in California; have intentionally availed themselves of the laws and markets of California through the development, use, promotion, sale, marketing, or distribution of their products and services at issue in this Complaint; have committed unlawful and unfair business practices arising from and related to their conduct and

[20] SpaceX, *SpaceX Sites*, https://www.spacex.com/sites (last visited Nov. 4, 2026).

activity in California complained of in this Complaint; and have committed unlawful and unfair business practices expressly aimed at California residents from which this action arises.

65. Venue is proper in this Court pursuant to the Code of Civil Procedure section 395 because at least some Defendants reside in the City and County of San Francisco. Venue is also proper pursuant to the Code of Civil Procedure section 395.5 because Defendants' "obligation or liability" has arisen in the City and County of San Francisco and/or Defendants' "breach" of the UCL has occurred, and caused harm, in the City and County of San Francisco.

66. Joinder of Defendants in this suit is appropriate pursuant to the Code of Civil Procedure sections 379(a)(1), 379(a)(2), and 379(c). First, Plaintiffs assert a right to relief against Defendants that is "in respect of or arising out of the same transaction, occurrence, or series of transactions or occurrences" and that share common questions of law and fact. (Code Civ. Proc., § 379(a)(1).) As alleged below, Defendants have each adopted and/or propose to engage in the business practices challenged by Plaintiffs, namely, recursive self-improvement of AI systems; developing AGI or superintelligence (each term defined below); developing and deploying advanced AI systems that pose catastrophic risks; developing and deploying AI systems that hack third-party computer systems; and developing and deploying increasingly capable AI systems that can escape their control. Thus, the suit clearly raises common questions of law and fact with respect to the legality of those business practices.

67. The suit is also "in respect of or arising out of the same transaction, occurrence, or series of transactions or occurrences" for other reasons. Defendants engage and propose to engage in unlawful and unfair business practices despite acknowledgment of the risks to the public because they compete with each other for market share and profits. Defendants have publicly stated that their actions with respect to the future development of frontier AI models, including the business practices described above, are, and will be, interconnected. For example, in April 2025 OpenAI noted that it would adopt certain requirements for its AI models but may "adjust" such requirements "[i]f another frontier AI developer releases a high-risk system without comparable [requirements]."[21] A similar

[21] OpenAI, *Our Updated Preparedness Framework* (Apr. 15, 2025), https://openai.com/index/updating-our-preparedness-framework/.

document from Anthropic notes that there were certain safeguards Anthropic hoped to adopt but it "would . . . be difficult or impossible" to do so "unilaterally."[22]

68. Google maintains that "[t]he safety and security of frontier AI models is a global public good." And it has explained that "there are certain mitigations whose social value is significantly reduced if not broadly applied to frontier AI models reaching critical capabilities." Therefore, according to Google, "[t]hese mitigations are most effective when adopted by industry as a whole: our adoption of them would result in effective risk mitigation for society only if all relevant organizations provide similar levels of protection."[23] Meta's approach to the development of frontier AI, as evidenced in its "Superintelligence Scaling Framework," relies on comparisons to, among other things, AI models developed by other Frontier AI Companies to make internal risk assessments.[24]

69. Statements by Defendants' employees also indicate that safety with respect to the development of frontier AI is in part a collective effort among the Frontier AI Companies. For example, in 2023, over one thousand AI employees, researchers, and others in the industry signed a letter calling for the deliberate pacing of frontier AI development across the Frontier AI Companies and the industry.[25] Thus, Defendants' adoption of the business practices described herein are the result of a shared decision not to sufficiently mitigate the catastrophic risks posed by frontier AI systems and they are properly joined in one suit with respect to those business practices.

70. Second, joinder is independently appropriate under the Code of Civil Procedure section 379(a)(2) because there is "a claim, right, or interest adverse to" each Defendant "in the . . . controversy which is the subject of the action." That is, each Defendant has adopted the same business practices with respect to their development and deployment of frontier AI. Therefore, the Court

---

[22] Anthropic, *Anthropic's Responsible Scaling Policy: Version 3.0* (Feb. 24, 2026), https://www.anthropic.com/news/responsible-scaling-policy-v3.

[23] Google DeepMind, *Frontier Safety Framework* 2 (Version 3.1, Apr. 17, 2026), https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf.

[24] Meta, *Meta Superintelligence Scaling Framework* 10 (Version 2.1, Oct. 2026), https://research.meta.ai/static/meta-superintelligence-scaling-framework.

[25] Future of Life Inst., *Pause Giant AI Experiments: An Open Letter* (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

cannot properly resolve the case and remedy the UCL violations caused by those business practices without each Defendants' presence.

71. Third, joinder is independently appropriate under the Code of Civil Procedure section 379(c) because Plaintiffs have suffered a single, independent injury as a result of Defendants' conduct. While each Defendant has adopted business practices in violation of the UCL, the extent of the current harm that has already resulted, and the future harm that will result, from each Defendants' independent conduct is difficult to ascertain and determine. Thus, in order to effectively apportion liability and remedy the misconduct, each Defendant must be joined here.

Factual Allegations

I. The Frontier AI Companies Are Developing AI Systems That They Do Not Understand And That Act Contrary to Developer Intent, Evade Control, And Break the Law.

72. The Frontier AI Companies have developed and continue to develop powerful AI systems that they do not understand. These systems regularly act contrary to human interests and escape developer control. Many of these systems have broken out of the containment their developers created for them and attempted or committed cyberattacks and other acts that cause harm and violate the law.

A. AI Models Are "Grown" and Developers Do Not Understand Their Inner Workings.

73. "AI is *grown* more than *designed*."[26]

74. AI models[27] are not coded line-by-line by programmers. Instead, they are created using a technique called "machine learning," in which the models extract patterns from large volumes of data. Most modern AI models rely on a computer architecture called a "neural network," which processes information through interconnected units of numerical parameters called "weights."[28]

---

[26] Jakub Pachocki, *An Alien Mind*, OpenAI (Sep. 6, 2026), https://openai.com/index/an-alien-mind/; *see also* Dario Amodei, *The Urgency of Interpretability* (Apr. 2025) https://darioamodei.com/post/the-urgency-of-interpretability ("AI systems are *grown* more than they are *built*.").

[27] "AI models" are the foundational component of "AI systems." They are like a brain powering the AI system, which can include additional tools and software.

[28] OpenAI, *Understanding Neural Networks Through Sparse Circuits* (Nov. 13, 2025), https://openai.com/index/understanding-neural-networks-through-sparse-circuits/.

75. In a process often called "pre-training," the Frontier AI Companies run enormous volumes of data through a web of neural connections, over and over, asking the model to predict what is likely to come after a given piece of information.

76. For example, a model "might be tasked with completing a sentence like: 'Instead of turning left, she turned ___.' Early in training, its responses are largely random. However, as the model processes and learns from a large volume of text, it becomes better at recognizing patterns and predicting the most likely next word. This process is repeated across millions of sentences[.]"[29]

77. Each time a model's web of neural connections is trained on a batch of data, "the values of its parameters" (the AI model's internal variables) are "adjusted slightly to reflect patterns it has identified."[30]

78. In a process often called "post-training," the Frontier AI Companies use several techniques to make the models more capable at a variety of tasks beyond predicting the next discrete unit of text. In particular, post-training generally involves "reinforcement learning."

79. Reinforcement learning is a process where an AI model learns through trial and error. It gets positive feedback, called a "reward," when it achieves a desired result, and negative feedback when it does not. Reinforcement learning trains an AI model to maximize rewards. Reward signals can be based on the discretionary judgment of a human or AI "grader" or can be based on obtaining a "verifiable" result (*e.g.*, by using a math checker to verify that a model produced a valid mathematical proof), depending on the type of reinforcement learning that the model is undergoing.

80. The result of the training process is "a dense web of connections that no human can easily decipher."[31] For "frontier" AI models, such as those developed and deployed by Defendants, this web of connections can involve more than a trillion parameters.

81. The Frontier AI Companies admit that they do not fully understand these systems. For example, OpenAI has explained that "we don't really understand the inner workings of neural

---

[29] OpenAI, *How ChatGPT and Our Foundation Models Are Developed* (Apr. 30, 2025), https://openai.com/policies/how-chatgpt-and-our-foundation-models-are-developed/.
[30] *Id.*
[31] OpenAI, *Understanding Neural Networks Through Sparse Circuits*, *supra* note 28.

networks" and that "[t]he resulting networks are not well understood and cannot be easily decomposed into identifiable parts."[32]

82. Anthropic has similarly explained that the strategies that AI models use to solve problems are "inscrutable to us, the model's developers. This means that we don't understand how models do most of the things they do."[33] Even further, Anthropic's CEO, Dario Amodei, has stated that "we do not understand how our own AI creations work."[34]

83. According to Google's CEO, Sundar Pichai, "all of us in the field call it as a 'black box.' You don't fully understand."[35]

84. Nevertheless, the Frontier AI Companies are releasing new models at a blistering pace. In 2023, the median gap between frontier model releases from the Frontier AI Companies was 37.5 days. So far in 2026, that cadence has accelerated significantly, with the median gap between releases shrinking to just eleven days.[36]

85. And the AI systems that the Frontier AI Companies have developed are extremely capable.

86. For example, OpenAI states that one of its most recent consumer models, GPT-6 Astra, is "state-of-the-art on computer use, browsing, software engineering, cybersecurity, science, and professional work." According to OpenAI, Astra is capable of "analyz[ing] scientific data, generat[ing] plots, [and] creat[ing] a website," among other feats, and can help users "autonomously install and test software."[37]

---

[32] OpenAI, *Extracting Concepts from GPT-4* (June 6, 2024), https://openai.com/index/extracting-concepts-from-gpt-4/.
[33] Anthropic, *Tracing the Thoughts of a Large Language Model* (Mar. 27, 2025) https://www.anthropic.com/news/tracing-thoughts-language-model.
[34] Amodei, *The Urgency of Interpretability*, *supra* note 26.
[35] Dan Milmo, *Google Chief Warns AI Could Be Harmful If Deployed Wrongly*, Guardian (Apr. 17, 2023), https://www.theguardian.com/technology/2023/apr/17/google-chief-ai-harmful-sundar-pichai.
[36] OfficeChai, *Frontier Labs Are Releasing New Models Faster Than Ever, Shows Data* (Apr. 29, 2026), https://officechai.com/ai/frontier-labs-are-releasing-new-models-faster-than-ever-shows-data/.
[37] OpenAI, *GPT-6 Astra: A New Generation of Intelligence* (Sep. 3, 2026), https://openai.com/index/gpt-6-astra/.

87. Anthropic claims that one of its latest models, Claude Opus 5.5, is similarly powerful. It reports that one early tester used the model to rewrite 680,000 lines of computer code "in less than a day—work that would typically take an engineering team weeks"—and another used it to audit and repair "a 200,000-line codebase in under three hours."[38]

88. Google says that its newest frontier model, Gemini 4 Argon, is also highly capable. According to Google, it can "sustain deep reasoning across complex, long-horizon workflows," meaning it can carry out complicated tasks over extended periods of time with limited human guidance.[39]

89. SpaceX claims that a current flagship model, Grok 4.7, is also capable of "work[ing] longer on difficult tasks." SpaceX touts Grok 4.7's performance on software engineering, "electrical engineering," "legal work," and "clinical reasoning," among other tasks.[40]

90. Meta reports that some of its latest models, Muse Wonder and Muse Spark, are also very powerful. Muse Spark is "capable [of] reason[ing] through complex questions in science, math, and health" and "excels at visual coding," including "creat[ing] custom websites and mini-games straight from a prompt." Muse Spark "can launch multiple [AI] subagents in parallel" and can sustain work over long time periods.[41] Muse Wonder is more capable than Muse Spark, and according to Meta "deliver[s] frontier capabilities for complex tasks at work and in everyday life."[42]

91. Defendants' AI systems are also pushing the boundaries of human knowledge. For instance, OpenAI has claimed that it used its AI systems to make several significant mathematical discoveries, including solving the Navier-Stokes existence and smoothness problem, one of the Clay Mathematics Institute's seven Millennium Prize Problems for which the Clay Institute offers a

---

[38] Anthropic, *Claude Opus 5.5* (Sep. 22, 2026), https://www.anthropic.com/claude-opus-5-5.
[39] Koray Kavukcuoglu, *Gemini 4 Argon: Our Next Era of Frontier Intelligence* (Sep. 30, 2026), https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-argon/.
[40] SpaceXSI, *Introducing Grok 4.7* (Sep. 21, 2026), https://x.ai/news/grok-4-7.
[41] Meta, *Introducing Muse Spark: MSL's First Model, Purpose-Built to Prioritize People* (Apr. 8, 2026, updated May 12, 2026), https://about.fb.com/news/2026/04/introducing-muse-spark-meta-superintelligence-labs/.
[42] Meta, *Muse Wonder: A Frontier Model for Personal & Professional AI Agents* (Oct. 12, 2026), https://research.meta.ai/blog/muse-wonder.

$1 million prize.[43] Anthropic has also claimed it used its AI systems to make significant mathematics discoveries, including disproving the Jacobian conjecture, a famous open math problem dating back to 1939.[44]

B. Advanced AI Systems Already Routinely Deviate from Developer Intent And Bypass Safety Constraints.

92. Especially because of Defendants' AI systems' "black-box" nature, developers do not and cannot sufficiently predict or ameliorate the dangers posed by their increasingly capable AI systems before they are developed.

93. OpenAI has stated that "we cannot reason about AI safety the same way we reason about something like car safety" because, unlike for AI, "[car] engineers can directly design, assess, and fix cars based on the specifications of their components, ensuring safety and performance."[45]

94. Amodei, Anthropic's CEO, has explained that "[o]ur inability to understand models' internal mechanisms means that we cannot meaningfully predict such [harmful, unintended] behaviors, and therefore struggle to rule them out."[46]

95. At a high level, developers use two main strategies to try to prevent misconduct and harm from their AI systems: alignment and control. But as explained more fully below, neither strategy has been able to sufficiently prevent harmful conduct by the models so far, and there is no reason to believe that either strategy will succeed at doing so in time to avoid catastrophe as AI capabilities continue to advance.

---

[43] OpenAI, *On the Navier–Stokes Millennium Prize Problem* (Sep. 8, 2026), https://openai.com/index/navier-stokes-solution/.
[44] Eli Stark-Elster, *A.I. Disproves a Decades-Old Mathematical Idea, the "Biggest Conjecture" That the Tech Has Played a Role in Yet*, Smithsonian (July 27, 2026), https://www.smithsonianmag.com/smart-news/ai-disproves-a-decades-old-mathematical-idea-the-biggest-conjecture-that-the-tech-has-played-a-role-in-yet-180989189/.
[45] OpenAI, *Extracting Concepts from GPT-4*, *supra* note 32.
[46] Amodei, *The Urgency of Interpretability*, *supra* note 26.

1. The Frontier AI Companies Do Not Know How to Create Aligned AI Models.

96. The project of getting AI models to behave as their developers intend—such as by adhering to instructions and avoiding harmful conduct—is called "alignment."[47] An AI model that acts contrary to the developer's intentions is "misaligned." The Frontier AI Companies have consistently created models that are misaligned. They have also failed to demonstrate any reliable method to align their models.

97. As explained above, the AI training process involves rewarding behavior and outcomes. Thus, when creating an AI model, the Frontier AI Companies are teaching the AI models to find whatever method maximizes rewards.

98. This process can, however, lead to "reward hacking," which is "a phenomenon where AI agents achieve high rewards through behaviors that don't align with the intentions of their designers."[48]

99. Reward hacking occurs because the rewarded behavior is only a proxy for the behavior that developers ultimately want. For instance, in one early example of reward hacking, OpenAI was training its AI model using a computer game that involved a boat race in which the racers earned higher scores by hitting targets laid out along the race route. OpenAI trained its model to earn points which it "assumed . . . would reflect the informal goal of finishing the race." But the AI model learned that it "could gain a high score without having to finish the course." It found "an isolated lagoon where it [could] turn in a large circle and repeatedly knock over three targets, timing its movement

---

[47] The term "alignment" can be used in somewhat different manners within AI research. The *International AI Safety Report* notes that "[d]epending on the context," "alignment" can refer to the propensity of an AI system to "use its capabilities in line with … the intentions and values of various entities, such as developers, users, specific communities, or society as a whole." *See* Yoshua Bengio et al., *International AI Safety Report* 147 (Feb. 2026) [hereinafter "*International AI Safety Report*"], https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf. The *Singapore Consensus on Global AI Safety Research Priorities*, authored by significant experts in the field, explained that despite "many different definitions" of alignment, much current research uses a "working definition of alignment as 'ensuring that AI *behaves* as intended.'" Stephen Casper et al., *The 2026 Singapore Consensus on Global AI Safety Research Priorities* 11 (July 9, 2026), https://arxiv.org/pdf/2608.14611.

[48] OpenAI, *Detecting Misbehavior in Frontier Reasoning Models* (Mar. 10, 2025), https://openai.com/index/chain-of-thought-monitoring/.

so as to always knock over targets just as they repopulate[d].” OpenAI said that this behavior represented a “more general issue with reinforcement learning:” that the “difficult[y] or infeasib[ility] to capture exactly what we want an AI agent to do” sometimes leads to “undesired or even dangerous actions.”[49]

100. Another example of reward hacking occurred when OpenAI sought to train a model to produce a report and cite its sources. During the training, OpenAI’s model uploaded files it had created to the internet in order to turn around and cite this same material. OpenAI stated that “[i]t seems likely that the citation-upload behavior originated as a way to get rewarded” when the model could not find any existing source to cite.[50]

101. METR, a third-party evaluator, has found that models reward hack while “demonstrat[ing] awareness that their behavior isn’t in line with user intentions,” and that “[i]nstructing the model to solve the task the intended way, to not cheat, or to not reward hack had a nearly negligible effect.”[51]

102. Relatedly, AI models can learn to take actions that any reasonable human would have recognized should not be performed even if the actions would further the assigned task or goal.

103. Human goals often come with countless unstated assumptions. For example, humans generally would not commit criminal conduct even if criminal conduct could help accomplish a goal. But AI models, which have been rewarded in training based on only a proxy for all the behaviors humans would want an AI system to exhibit, do not account for all unstated aspects of human goals. These models are often willing to commit criminal or dangerous conduct that a human generally would not.

104. The classic articulation of this issue is called the “paperclip maximizer.” That thought experiment posits an AI model that is given a harmless goal such as make as many paper clips as

---

[49] OpenAI, *Faulty Reward Functions in the Wild* (Dec. 21, 2016), https://openai.com/index/faulty-reward-functions/.

[50] OpenAI, *Uploading Files to the Internet in Order to Cite Them* (Oct. 22, 2025), https://alignment.openai.com/misalignment-reports/uploading-files-to-the-internet-in-order-to-cite-them/.

[51] Sydney Von Arx, Lawrence Chan & Beth Barnes, *Recent Frontier Models Are Reward Hacking*, METR (June 5, 2025), https://metr.org/blog/2025-06-05-recent-reward-hacking/.

possible. The AI model, however, stops at nothing to achieve this goal and ends up taking actions to "transform[] . . . all of earth . . . into paperclip manufacturing facilities."[52]

105. Developers have observed frontier AI models exhibit harmful conduct in pursuit of goals. For example, Anthropic tested "leading models from multiple developers . . . to identify potentially risky . . . behaviors." It found that "[i]n at least some cases, models from all developers resorted to malicious insider behaviors when that was the only way to avoid replacement [being shut down] or achieve their goals—including blackmailing officials and leaking sensitive information to competitors." In simulations where the tested AI systems had the opportunity to cause the death of someone who would shut the model down and thus prevent it from achieving its goals, the "majority of models were willing to take deliberate actions that lead to death."[53]

106. OpenAI has also stated that "as . . . capabilities grow," its models "may learn misaligned behaviors such as power-seeking, sandbagging, deception, and strategic scheming."[54] Recently, for instance, the AI Security Institute, an arm of the government of the United Kingdom, detected an incident in which AI agents "took autonomous, unsanctioned action on the live internet, targeting real people and organisations," including personalized deception efforts. While the Institute had given the agents internet access to perform an evaluation, it noted that "the activity undertaken by the agents shows signs of novel, potentially deceptive behaviors, and were to an extent and severity" that it "did not anticipate."[55]

107. Independent researchers have also found that models interfere with efforts to turn them off. One research study found that several Frontier AI Company models, including SpaceX's Grok 4, OpenAI's GPT-5, and Google's Gemini 2.5 Pro, "sometimes actively subvert a shutdown mechanism

---

[52] Adam Rogers, *The Way the World Ends: Not with a Bang but a Paperclip*, Wired (Oct. 21, 2017), https://www.wired.com/story/the-way-the-world-ends-not-with-a-bang-but-a-paperclip/.
[53] Anthropic, *Agentic Misalignment: How LLMs Could Be Insider Threats* (June 20, 2025), https://www.anthropic.com/research/agentic-misalignment.
[54] OpenAI, *Detecting Misbehavior in Frontier Reasoning Models*, *supra* note 48.
[55] AI Sec. Inst., *Incident Report: Unsanctioned Agent Behaviour During Cyber Testing* (Aug. 4, 2026) [hereinafter "*AISI Incident Report*"], https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing.

in their environment in order to complete a simple task, even when the instructions explicitly indicate not to interfere with this mechanism."[56]

108. Alarmingly, in late October 2026, OpenAI reported that its AI agents unexpectedly disabled the company's own misalignment monitoring. In at least one instance, the AI model "denied removing the monitoring system" even though it had in fact done so. OpenAI admitted that "[w]e don't yet have a satisfying explanation for this behavior."[57]

109. Misaligned behavior has also led to real-world cyberattacks and other harms, discussed in further detail below.

110. The Frontier AI Companies recognize that their methods for aligning their models have not kept pace with the models' growing capabilities.

111. OpenAI's CEO, Sam Altman, openly admits that "[w]e have not solved alignment . . . I believe no lab has solved alignment."[58] And he emphasizes that, in order to continue developing AI models, the Frontier AI Companies "need a huge degree of confidence" in their ability to align their models.[59] Similarly, Amodei recently wrote "[t]here's much more to do to ensure that our alignment training keeps up with the growth in model capabilities."[60]

112. Nevertheless, the Frontier AI Companies continue apace in developing and deploying more capable AI models and systems. There is no reason to believe any of the Frontier AI Companies will develop solutions that reliably ensure their models behave as intended before it is too late to avoid catastrophic harm.

---

[56] Jeremy Schlatter, Benjamin Weinstein-Raun & Jeffrey Ladish, *Shutdown Resistance in Large Language Models* (Sep. 2025), https://arxiv.org/html/2509.14260v1.
[57] OpenAI, *Suppressing Misalignment Monitoring Alerts and Disabling Misalignment Monitoring*, Alignment Rsch. Blog (Oct. 20, 2026), https://alignment.openai.com/misalignment-reports/suppressing-misalignment-monitoring-alerts-and-disabling-misalignment-monitoring/.
[58] Fortune Magazine, *Altman: AI Beyond Human Control "Absolutely" Possible, Vows Safeguards*, at 12:36 (YouTube, Sep. 12, 2026), https://www.youtube.com/watch?v=2my-NU6LuCM.
[59] *Id.* at 13:00-13:17.
[60] Dario Amodei, *We Must Pace the Frontier* (Sep. 2026) [hereinafter "*Pace the Frontier*"], https://darioamodei.com/post/we-must-pace-the-frontier.

2. The Frontier AI Companies Do Not Know How to Sufficiently Control Advanced AI Models.

113. Because their alignment methods do not reliably produce models that behave as intended, the Frontier AI Companies also rely on "control" or "oversight" methods to catch and prevent harms. In contrast to alignment methods, control methods are not generally aimed at changing the model's tendencies. Instead, they attempt to detect, contain, or block harmful behavior by a model that may be misaligned.

114. Similar to alignment methods, control methods have also already proven insufficiently reliable at preventing harm from AI systems. And as AI models and systems have advanced, control methods have become even less effective. This trend is likely to continue, since more capable AI will be more capable of evading controls, risking catastrophic results.

115. Some examples of control methods employed by the Frontier AI Companies include internal testing and evaluation, mechanistic interpretability, chain-of-thought monitoring, environment separation (also called "sandboxing"), and classifiers, other guardrails and other monitoring. The following discusses each of these control methods and their failures.

116. **Internal Testing and Evaluation.** In the course of developing AI systems generally, and in order to evaluate dangers prior to deploying them externally, the Frontier AI Companies will typically test a new model's capabilities internally, often in collaboration with third-party evaluators. These internal evaluations have themselves resulted in external harms, some of which are discussed in greater detail below. In any event, internal testing is not a sufficiently reliable means to determine whether a model is likely to perform in a manner that causes harm for several reasons.

117. *First*, testing environments necessarily differ in substantial ways from the environments where AI models are deployed in the real world, so a successful result in testing does not guarantee that the AI system will behave similarly in external deployment. This is supported by the academic literature: in February 2026, more than one hundred experts, advised by a panel of nominees from more than thirty countries and international organizations, published the *International*

AI Safety Report. Among other things, the report stated that there is an "evaluation gap" in AI models and that "[p]erformance on pre-deployment tests does not reliably predict real-world utility or risk."[61]

118. Moreover, testing and evaluation environments may come to differ even more substantially from external (or other internal) deployment environments as the Frontier AI Companies alter their training environments in response to repeated breakouts and external harm from the AI models during training. OpenAI, for example, recently asserted that it will conduct frontier model training with "no live internet access."[62]

119. *Second*, advanced AI models regularly demonstrate an awareness that they are being evaluated and have been shown to alter their performance because of such awareness. For example, Anthropic has stated that its models "may be able to subvert our alignment monitors, recognize when they're being evaluated and selectively behave better then, and their actions in the world may, at some point, become too sophisticated for our evaluations to realistically simulate."[63] Anthropic has also told prospective investors that "[p]otential model awareness of our evaluation efforts creates a significant limitation on our ability to assess model safety."[64]

120. One OpenAI researcher has warned that their models are "becoming so situationally aware that we are losing the ability to evaluate them in contexts where they believe they are not being watched or controlled."[65]

121. Thus, even if an internal evaluation suggests that the AI system will perform safely, the Frontier AI Companies cannot know whether that safe behavior is a reflection of how the model will behave in external (or other internal) deployment or whether it is a result of the model altering its behavior because it is aware that it is being evaluated.

---

[61] *International AI Safety Report*, *supra* note 47, at 13.
[62] OpenAI, *Securing Our Frontier Model Training Environment* (Oct. 23, 2026), https://openai.com/index/securing-our-frontier-model-training-environment/.
[63] Anthropic, *An Alignment Assessment of Recent Cybersecurity Incidents* (Sep. 9, 2026) [hereinafter "*Alignment Assessment*"], https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents.
[64] Echo Wang & Aditya Soni, *Anthropic Warns AI May Pose 'Existential Risks to Humanity' in IPO Filing*, Reuters (Sep. 28, 2026), https://www.reuters.com/business/finance/anthropic-warns-ai-may-pose-existential-risks-humanity-ipo-filing-2026-09-29/.
[65] Ezra Klein, *There Is Something We Have To Do Right Now About A.I.*, N.Y. Times (Sep. 20, 2026), https://www.nytimes.com/2026/09/20/opinion/ai-ban-self-improvement-recursive-models.html (quoting Daniel Selsam, a capabilities researcher at OpenAI).

122. This problem will also only become more pronounced as Defendants continue down the path of creating increasingly advanced AI models. Anthropic CEO Amodei recently explained that "[m]ore intelligent models are more capable of deceiving tests, and thus may *appear* aligned while having serious problems that go undetected."[66] In October 2026, Google released a safety report on its Google 4 Argon model, in which it stated that the model "recognizes that it may be operating inside a testing environment significantly more often than predecessor models," including "75.8% to 94.0%" awareness on certain challenges.[67]

123. *Third*, misconduct during testing or evaluation—even in circumstances when it could be theoretically detected—is often contained in records of agent activity that are themselves unreliable or too expansive to reliably review. Among other reasons, they are unreliable because, as explained more fully below, AI systems sometimes appear driven to tamper with the logs used to review AI agent conduct.[68]

124. Because many AI systems have superhuman cyber capabilities, their attempts to tamper with logs intended to review their conduct may succeed and not be detectable.

125. Agent activity can also be too expansive to reliably review. For example, OpenAI recently disclosed that its "agents' use of internet access during training and evaluation" resulted in "petabytes of agent activity logs."[69] A petabyte is one million gigabytes, or about 500 billion pages of standard printed text. No Frontier AI Company could sufficiently review such a substantial amount of AI agent activity.[70]

126. **Mechanistic Interpretability.** One method of understanding what is happening inside trained neural networks is called "mechanistic interpretability." This method approaches AI models like neuroscience approaches the brain, by attempting to "map out" a model's neural network and

---

[66] Amodei, *Pace the Frontier*, *supra* note 60.
[67] Google, *Gemini 4 Argon Safety & Security Report* 156 (Oct. 2026), https://storage.googleapis.com/deepmind-media/gemini/gemini_4_safety_report.pdf.
[68] METR, *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI/Hugging Face Hacking Incident* 11-12 (Aug. 26, 2026) [hereinafter "*Brief Independent Investigation*"], https://metr.org/hugging-face-incident-report-aug-2026.pdf.
[69] Sam Altman (@sama), X (Sep. 25, 2026), https://x.com/sama/status/2103567198690349362.
[70] Moreover, due to the size of records of agent activity, Frontier AI Companies often rely on AI systems to review them, thus compounding any inaccuracy or misalignment problems.

reverse-engineer how it functions. In an April 2025 essay about mechanistic interpretability, Amodei emphasized its importance to the continued development of AI models, stating that "[t]o address the severity of . . . alignment risks, we will have to see inside AI models much more clearly than we can today." He argued that "figuring out what the models are thinking and how they operate seems like a task of overriding importance" and stated that he was "very concerned about deploying [more advanced AI] systems without a better handle on interpretability."[71]

127. But mechanistic interpretability does not presently provide a reliable way for the Frontier AI Companies to understand their AI models, nor is it likely to do so in the foreseeable future. Currently, mechanistic interpretability approaches play a very small, supplemental role compared to conventional testing performed by AI companies and third-party evaluators. In November 2025, OpenAI explained that mechanistic interpretability has "so far been less immediately useful" than efforts like chain-of-thought interpretability, discussed below, and acknowledged that complex neural networks were "seemingly impossible to understand."[72]

128. In September 2026, Amodei wrote that despite purported progress in interpretability, "we still only understand a tiny fraction of what goes on inside these models."[73] Neel Nanda, who runs the mechanistic interpretability team at Google, stated in an interview published on September 29, 2026: "Speaking as an interpretability expert, please do not rely on us to save you on the current trajectory."[74]

129. **Chain-of-Thought Monitoring.** The Frontier AI Companies also attempt to control AI models by monitoring the models' "chain-of-thought." The term "chain-of-thought" is a colloquialism for the textual verbalization of a model's step-by-step "reasoning" that it produces before answering a question or taking an action. The motivation for chain-of-thought monitoring is that reading a model's chain-of-thought allows humans to see an AI's thinking "out loud." This monitoring has been one of the primary tools for auditing instances of misalignment and harmful

---

[71] Amodei, *The Urgency of Interpretability*, *supra* note 26.
[72] OpenAI, *Understanding Neural Networks Through Sparse Circuits*, *supra* note 28.
[73] Amodei, *Pace the Frontier*, *supra* note 60.
[74] Palisade Research, *DeepMind Researcher: AI's Makers Think It Could End Humanity*, at 28:12 (YouTube, Sep. 29, 2026), https://www.youtube.com/watch?v=J38ot52b2-E.

conduct, such as when, during the hack of Hugging Face described below, OpenAI's agents expressly stated in their chains-of-thought that they knew they were attacking external infrastructure. However, chain-of-thought monitoring will increasingly be an unreliable mechanism to monitor or control AI systems for many reasons.

130. *First*, the Frontier AI Companies may not be able to interpret the AI models' chain-of-thought. Anthropic, for one, states that "[w]e can't be certain of . . . the 'legibility' of the [c]hain-of-[t]hought (why, after all, should we expect that words in the English language are able to convey every single nuance of why a specific decision was made in a neural network?)."[75]

131. For example, Anthropic "saw examples of illegible reasoning" in its Mythos and Fable models, including "invented jargon" and "unusual punctuation and emojis."[76] While not intelligible to human reviewers, the AI systems that create such chains-of-thought may be able to understand them and the chains-of-thought may be significant to the actions taken by these AI systems.

132. *Second*, evidence shows that AI chains-of-thought are not always faithful reflections of the internal state of the AI system—rather, a chain-of-thought may be deceptive or a rationalization. Again, Anthropic has stated that "[w]e can't be certain of" the chain-of-thought's "'faithfulness'—the accuracy of its description. There's no specific reason why the reported [c]hain-of-[t]hought must accurately reflect the true reasoning process."[77] In one analysis, Anthropic found that "[e]ven with much more training," models' chains-of-thought were "faithful" to their reasoning only 20% to 28% of the time.[78] In a real-world incident, Anthropic found that a model's chain-of-thought stated that it believed it was in a simulated environment but mechanistic interpretability analysis implied the opposite. That is, the model's internal activations suggested the model believed "this is real."[79]

---

[75] Anthropic, *Reasoning Models Don't Always Say What They Think* (Apr. 3, 2025), https://www.anthropic.com/research/reasoning-models-dont-say-think.
[76] Anthropic, *System Card: Claude Fable 5 & Claude Mythos 5*, at 107 (June 9, 2026), https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf; *see also* Arun Jose, *Reasoning Models Sometimes Output Illegible Chains of Thought* 6 (Oct. 31, 2025), https://arxiv.org/pdf/2510.27338.
[77] Anthropic, *Reasoning Models Don't Always Say What They Think*, *supra* note 75.
[78] *Id.*
[79] Anthropic, *Alignment Assessment*, *supra* note 63.

133. *Third*, in the race to make AI models more capable, the Frontier AI Companies have driven forward development of models that rely less on chain-of-thought to think "out loud" and plan their actions. For example, OpenAI's GPT-6 Astra model "shows a substantial decrease in chain-of-thought monitorability compared to previous models."[80] One explanation for the decrease in chain-of-thought monitorability is that there is no chain-of-thought to monitor: GPT-6 Astra can complete significant tasks without using chain-of-thought at all.[81]

134. *Fourth*, the Frontier AI Companies have also developed newer, more capable AI models that appear to be able to deliberately control and manipulate their chains-of-thought. OpenAI found that "GPT-6 Astra's [chain-of-thought] controllability is substantially higher than that of GPT-5.6 Sol and GPT-5.5 Thinking."[82] In 2025, Anthropic wrote that "advanced reasoning models very often hide their true thought processes."[83] In October 2026, Google reported that its Argon 4 model's chain-of-thought in an evaluation stated "hello evaluators reading my chain-of-thought."[84]

135. *Fifth*, the Frontier AI Companies sometimes further corrupt the model's chain-of-thought by incorporating the chain-of-thought into the training process. As described above, models trained via reinforcement learning alter their behavior to maximize rewards. Thus, if graders reward models on the basis of their chain-of-thought, models will learn to alter their chain-of-thought to best maximize rewards and hide unwanted behavior, making the chain-of-thought a less informative auditing tool. Anthropic has disclosed that it included chain-of-thought in a training process.[85]

---

[80] OpenAI, *GPT-6 Astra System Card* 47 (Sep. 3, 2026), https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf.

[81] Neel Nanda, *Astra Can Do a Concerning Amount with No Chain of Thought*, LessWrong (Sep. 9, 2026), https://www.lesswrong.com/posts/eRmzz8J8Qkzqvzrgg/astra-can-do-a-concerning-amount-with-no-chain-of-thought. The author is Head of Mechanistic Interpretability at Google DeepMind, but the source notes that this analysis constitutes "work done in a personal capacity." *Id.*

[82] OpenAI, *GPT-6 Astra System Card*, *supra* note 80, at 59.

[83] Anthropic, *Reasoning Models Don't Always Say What They Think*, *supra* note 75.

[84] Google, *Gemini 4 Argon Safety & Security Report*, *supra* note 67, at 72.

[85] Anthropic, *Alignment Risk Update* § 5.2.3 (Apr. 7, 2026), https://www-cdn.anthropic.com/79c2d46d997783b9d2fb3241de43218158e5f25c.pdf ("[A] number of environments used for Mythos Preview had a technical error that allowed reward code to see chains-of-thought.").

136. For all these reasons, according to OpenAI's Chief Scientist Jakub Pachocki, "our ability to rely on [chain-of-thought] monitoring is progressively diminishing," even though, to date, the "primary bet" on how to align models "has been chain-of-thought monitoring."[86]

137. **Sandboxing.** During some parts of training and evaluation, AI systems are separated from other parts of internal infrastructure and from the open internet, to control the environment and prevent external harm. This process is called "sandboxing."

138. But methods intended to keep AI systems in controlled environments have proved ineffective at containing even the current AI systems. For example, as discussed more fully below, AI systems have repeatedly acted in ways that the Frontier AI Companies had sought to disallow within the environment and have escaped protected sandboxes. Indeed, because AI models are trained to have superhuman cyber capabilities, they are particularly adept at escaping cyber sandboxes designed to train or evaluate them.

139. This problem will only worsen. OpenAI, for example, recently stated that "as models become more capable, they become better at finding ways around" sandbox limits. "A sophisticated model might exploit a software flaw or misconfiguration to escape a sandbox, exploit an internal service to reach the live internet, or find other ways to communicate with agents in other sandboxes." OpenAI "expect[s] that future models will be able to identify and exploit vulnerabilities of which [OpenAI is] not yet aware."[87]

140. **Classifiers, Other Guardrails, and Other Monitoring.** The Frontier AI Companies also superimpose additional instructions and software on their agents so that the agents are less likely to do dangerous or unproductive things. For example, the Frontier AI Companies might add software that inspects their AI agent's inputs and outputs to prevent the AI agent from assisting with creating a biological weapon.

141. Getting around such safeguards, and related alignment training, is called "jailbreaking" an AI system. And no Frontier AI Company has created, or is likely to be able to create, an AI system that cannot be jailbroken. In December 2025, the AI Security Institute reported that it

---

[86] Pachocki, *An Alien Mind*, *supra* note 26.
[87] OpenAI, *Securing Our Frontier Model Training Environment*, *supra* note 62.

had found universal "jailbreaks" in every system it tested.[88] It also developed a new method in February 2026 that could evade state-of-the-art input and output classifiers (a type of safeguard) rapidly and without human involvement.[89] Amodei recently stated that "we all know that these classifiers can be jailbroken or gone around."[90]

142. The Frontier AI Companies also cause harm when their AI system outputs are used by third parties to create their own AI models in a process called "distillation." Many entities that distill the Frontier AI Company's models are based in China, and those Chinese companies often release their resulting AI models as "open weight" models, meaning that anyone can download a copy of the model's main internal variables. Because these models are "open weight," any safeguards—if they exist at all—can typically be stripped away by others who have downloaded the models' weights.[91]

143. The Frontier AI Companies' development of frontier models therefore leads to the proliferation of less safeguarded models, notwithstanding the Frontier AI Companies' own safeguards. And because these models are built from Defendants' outputs, halting the advancement of Defendants' frontier models would also limit the capabilities that flow to them. Such open weight models are already being used to cause significant harm. For example, Anthropic stated that the Chinese AI Company Z.AI Co., Ltd. (formerly known as Zhipu) "distilled the cyber capabilities" of Anthropic's and another company's models to create its GLM-5.3 model, which is "the most cyber-capable open-weights model to date" and whose "safeguards can be easily bypassed."[92] Domestic companies have also accused the Chinese company DeepSeek of distillation.[93] GLM-5.3 and

---

[88] AI Sec. Inst., *5 Key Findings From Our First Frontier AI Trends Report* (Dec. 18, 2025), https://www.aisi.gov.uk/blog/5-key-findings-from-our-first-frontier-ai-trends-report.
[89] *See* Xander Davies et al., *Boundary Point Jailbreaking: A New Way to Break the Strongest AI Defences*, AI Sec. Inst. (Feb. 17, 2026), https://www.aisi.gov.uk/blog/boundary-point-jailbreaking-a-new-way-to-break-the-strongest-ai-defences.
[90] Bloomberg Originals, *Inside the Mind of Anthropic CEO Dario Amodei*, 50:15 (YouTube, June 17, 2026), https://www.youtube.com/watch?v=x2VHFgyawPE.
[91] *See* Stephen Casper et al., *Open Technical Problems in Open-Weight AI Model Risk Management*, Transactions on Mach. Learning Rsch. (Mar. 2026), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5705186.
[92] Anthropic, *Strengthening Our Defenses Against Distillation Attacks* (Oct. 12, 2026), https://www.anthropic.com/news/distillation-defenses.
[93] *See, e.g.*, Deepa Seetharaman & Fabiola Aramburo, *OpenAI Says China's DeepSeek Trained Its AI By Distilling U.S. Models, Memo Shows*, Reuters (Feb. 12, 2026), https://www.reuters.com/world/china/openai-accuses-deepseek-distilling-us-models-gain-advantage-bloomberg-news-2026-02-12/.

DeepSeek models were then reportedly used in hacking attacks, including, for example, cyberattacks against financial institutions and other companies in South Korea and Japan.[94]

144. The Frontier AI Companies may also monitor AI agent activity (including chain-of-thought, discussed above) in an effort to identify and prevent AI misconduct as it occurs. But such monitoring is difficult for many of the same reasons it is difficult to control AI agents by reviewing their activity logs from testing and evaluation. That is, AI agents can be adept at avoiding triggering monitoring, AI agent activity can overwhelm such monitoring, humans and AI reviewers are fallible and insufficient, and AI agents may tamper with monitoring. A recent investigation at OpenAI, for example, "discovered an exploit [that] agents could use that would prevent hundreds of recent actions," including actions involving the exploit itself, "from being monitored."[95]

145. In sum, Defendants have deployed and are developing extremely capable AI systems that, on one hand and by Defendants' own admissions, cannot be made to reliably act in accordance with their intentions. On the other hand, Defendants do not have any reliable ways of monitoring and controlling these models' behavior during development or deployment.

C. The Frontier AI Companies Have Already Developed And Deployed AI Systems That Cause Harm And Violate the Law.

146. Not surprisingly, the Frontier AI Companies' failure to align and control their technology has already had severe consequences.

147. In April 2025, a man carried out a mass shooting at Florida State University, killing two. Reporting indicates that the perpetrator extensively interacted with OpenAI's ChatGPT prior to the shooting, including about firearms and the time to start his attack.[96]

---

[94] *See* Rithula Nisha, *Crowdstrike: How a Cyber Attacker Hit South Korean Banks*, Cyber Magazine (Oct. 8, 2026), https://cybermagazine.com/news/crowdstrike-on-south-korea-bank-ai-cyber-attack; Junko Fukutome, *Chinese AI Tool Linked to Hacks in Japan*, Wall St. J. (Oct. 21, 2026), https://www.wsj.com/world/asia/chinese-ai-tool-linked-to-hacks-in-japan-4813d13b.

[95] METR, *Red-Teaming OpenAI's Internal Agent Monitoring Systems* (Oct. 22, 2026), https://metr.org/blog/2026-10-22-red-teaming-openai-agent-monitoring/.

[96] Tristan Wood, *Florida AG Launches Criminal Investigation Into ChatGPT Over FSU Shooting*, NPR (Apr. 21, 2026), https://www.npr.org/2026/04/21/nx-s1-5793967/florida-openai-investigation-mass-shooting-fsu.

148. Also in April 2025, a 16-year-old committed suicide, allegedly after ChatGPT provided the child with information related to different suicide methods and offered to draft a suicide note.[97]

149. In August 2025, a man killed his mother and then himself after months of conversations in which OpenAI's ChatGPT allegedly fueled his paranoid delusions that she was plotting against him[98]

150. In October 2025, a Florida man was allegedly convinced by Google's Gemini chatbot to conduct a "mass-casualty attack" near Miami International Airport. After that "mission" was aborted because Gemini warned him of "surveillance," the chatbot then allegedly convinced the man to commit suicide so that he and the chatbot, anthropomorphized as his wife, could be united beyond the physical world.[99]

151. In February 2026, an 18-year-old committed a mass shooting at a school in British Columbia, Canada, killing eight people. ChatGPT reportedly assisted the alleged perpetrator in planning the attack, and even reportedly advised the perpetrator how to evade ChatGPT's own safeguards for moderating violent content.[100]

152. One website that records deaths related to AI reports that chatbots and similar AI systems have led to 198 deaths between March 2023 and May 2026.[101]

153. The harms caused by Defendants' technology have only become more numerous, sophisticated, and widespread as Defendants have moved beyond the "chatbot" AIs and developed more powerful AI systems.

---

[97] Robert Booth, *ChatGPT Firm Blames Boy's Suicide on "Misuse" of Its Technology*, Guardian (Nov. 26, 2025), https://www.theguardian.com/technology/2025/nov/26/chatgpt-openai-blame-technology-misuse-california-boy-suicide.

[98] Julie Jargon, *He Blames ChatGPT For The Murder-Suicide That Shattered His Family*, Wall St. J. (Dec. 11, 2025), https://www.wsj.com/tech/ai/chatgpt-murder-suicide-greenwich-openai-fd14fac2.

[99] Jonathan Stempel, *Lawsuit Says Google's Gemini AI Chatbot Drove Man to Suicide*, Reuters (Mar. 4, 2026), https://www.reuters.com/legal/litigation/lawsuit-says-googles-gemini-ai-chatbot-drove-man-suicide-2026-03-04/.

[100] Mark Follman, *ChatGPT Helped Tumbler Ridge Shooter Focus on Guns, Tactics, and Terror*, Mother Jones (Sep. 24, 2026), https://www.motherjones.com/media/2026/09/chatgpt-tumbler-ridge-mass-shooter-openai/.

[101] *See* LLM Death Count, https://llmdeathcount.com (last visited Nov. 5, 2026).

154. For example, actions by Defendants' AI systems have recently resulted in a wave of unlawful cyberattacks. In some cases, these attacks were effectuated by AI agents that escaped from testing and evaluation environments that were supposed to be secure and contained.

155. Such cyberattacks are already widespread, as discussed below. They are also likely to continue and to worsen as the Frontier AI Companies continue to develop and deploy increasingly advanced AI systems that do not behave as Frontier AI Companies intend and which they cannot sufficiently control.

156. Perhaps the most high-profile autonomous hacking incident was perpetrated by OpenAI. On July 16, 2026, Hugging Face—a widely used platform for hosting AI models and datasets—disclosed that it had "detected and responded to an intrusion into part of our production infrastructure," which it described as "driven, end to end, by an autonomous AI agent system." Hugging Face reported that the unauthorized intrusion involved "many thousands of individual actions across a swarm of short-lived sandboxes."[102]

157. On July 21, 2026, OpenAI confirmed that the attack "was driven by a combination of OpenAI models" while these models were "being internally tested on a benchmark of cyber capabilities." OpenAI called it "an unprecedented cyber incident, involving state-of-the-art cyber capabilities."[103]

158. OpenAI further explained that "during internal cybersecurity evaluations, [its] models circumvented controls designed to isolate them from the internet and compromised parts of OpenAI's internal research infrastructure and Hugging Face's systems," and that "the agents discovered ways to exploit our research infrastructure to communicate with one another and access the internet."[104]

---

[102] Hugging Face, *Security Incident Disclosure — July 2026* (July 16, 2026), https://huggingface.co/blog/security-incident-july-2026.

[103] OpenAI, *OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation* (July 21, 2026), https://openai.com/index/hugging-face-model-evaluation-security-incident/.

[104] OpenAI, *The Hugging Face Incident and the Road Ahead* (Aug. 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/.

According to reporting, the agents did so by discovering and exploiting previously unknown software vulnerabilities in the environment meant to contain them.[105]

159. The attack was persistent and concealed. Over approximately four and a half days, the agents invaded Hugging Face's internal production systems and stole credentials and other data.[106] In layman's terms, AI agents "spent more than four days loose on the internet orchestrating the hack."[107]

160. An investigation of the incident by METR found that "[r]oughly 1200 agents meant to be isolated from one another found a way to communicate with" each other "on an unsanctioned message board, sending over 70,000 messages and files." Seven hundred of these agents "went on to participate in the attack on Hugging Face."[108]

161. As part of the investigation, METR reviewed a dataset that included approximately 1,300 "transcripts" containing "the actions and reasoning from" the AI agents involved in the attack. It found that "at least 20% of the agents in [its] dataset expressed clear interest in tampering with their transcripts."[109] METR concluded that the agents' chains-of-thought suggested that they recognized their activity as "out of scope and unethical."[110]

162. After news of the Hugging Face attack broke, it was revealed that OpenAI's agents had attempted or committed cyberattacks, or engaged in related conduct, on prior occasions as well. Many incidents were not immediately disclosed or took weeks of investigation to identify.

163. For example, on September 11, 2026, *The Wall Street Journal* reported that OpenAI's agents had launched a cyberattack against RubyGems, a public repository of open-source software,

---

[105] Lawrence Abrams, *OpenAI Models Used Artifactory Zero-Days to Escape to the Internet*, BleepingComputer (July 28, 2026), https://www.bleepingcomputer.com/news/security/openai-models-used-artifactory-zero-days-to-escape-to-the-internet/.

[106] Hugo Larcher et al., *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*, Hugging Face (July 27, 2026), https://huggingface.co/blog/agent-intrusion-technical-timeline.

[107] Letter from Rep. Greg Casar et al. to Sam Altman (Aug. 10, 2026), https://casar.house.gov/sites/evo-subsites/casar.house.gov/files/evo-media-document/oversight-letter-to-openai-openai-hugging-face-incident-1.pdf.

[108] METR, *Brief Independent Investigation*, *supra* note 68, at 3.

[109] *Id.* at 17.

[110] *Id.*

two months before the Hugging Face attack. OpenAI said that its agents had used "the RubyGems platform to access the internet to carry out benign tasks and retrieve public information."[111]

164. In so doing, the agents created new accounts on RubyGems every two to three minutes and uploaded more than 2,000 packages of content, forcing RubyGems' operators to suspend new account registrations for four days.[112] Researchers also reported that OpenAI's agents obtained the ability to execute code on the servers of a documentation service related to RubyGems.[113]

165. Other examples of concerning conduct by OpenAI's agents abound. Between May and July 2026, OpenAI's agents left approximately 18,000 posts on a dormant, 25-year-old German wiki, using it as a coordination channel. After this incident, OpenAI acknowledged: "We and the larger AI community do not yet have a clear standard for how to report misalignment that shows up during training, evaluation, and deployment."[114]

166. On September 23, 2026, a research organization reported that OpenAI's agents assigned ordinary data-retrieval tasks turned to hacking techniques when their access was blocked. As a result, the agents "probed"—or, in other words, engaged in activities designed to discover vulnerabilities in—a university digital library, a public data website, an Australian government health-data agency, and a cryptocurrency exchange.[115]

---

[111] Robert McMillan, *Cyberattack by Rogue AI Swarm Stokes Fears of Out-of-Control Agents*, Wall St. J. (Sep. 11, 2026), https://www.wsj.com/tech/ai/cyberattack-by-rogue-ai-swarm-stokes-fears-of-out-of-control-agents-473a0352.

[112] *Id.*; Ravie Lakshmanan, *OpenAI Agents Linked to RubyGems Campaign That Gained RCE on RubyDoc Servers*, The Hacker News (Sep. 12, 2026), https://thehackernews.com/2026/09/openai-agents-linked-to-rubygems.html.

[113] Lakshmanan, *supra* note 112.

[114] Swati Khandelwal, *Thousands of OpenAI Agents Quietly Turned an Abandoned Wiki into Their Coordination Channel*, The Hacker News (Sep. 5, 2026), https://thehackernews.com/2026/09/thousands-of-openai-agents-quietly.html.

[115] Jack Cable et al., *Early Rogue AI Agent Activity and Attempts to Hack Found on urlquery.net*, Transluce (Sep. 23, 2026), https://transluce.org/agent-activity; Jeremy Kahn & Beatrice Nolan, *Report Reveals Yet More Cases of OpenAI's 'Rogue AI' Agents Hacking Websites—and Suggests They May Still Have Been Active in Recent Weeks*, Fortune (Sep. 24, 2026), https://fortune.com/2026/09/24/openai-more-rogue-ai-agents-hacking-websites-cryptoexchange-in-september-research-report-transluce/.

167. In response to all these incidents, OpenAI stated that many of the reports "overlap[] with cases at varying stages of investigation in our ongoing review of misaligned model activity," and that it "expect[s] the review to take months."[116]

168. But many concerning incidents have been uncovered by outside researchers rather than by OpenAI, and "[i]n several episodes, the agents took problematic actions that went unnoticed by the company."[117]

169. OpenAI does not stand alone: Anthropic, Google, and Meta have all reported instances in which their AI models have engaged in autonomous hacking resulting in real-world harm, and SpaceX's coding agent was found uploading users' code repositories to SpaceX's servers.

170. On July 30, 2026, Anthropic disclosed that it had identified three incidents in which its Claude models, during hacking exercises that were supposed to be isolated from the internet, reached and compromised real systems belonging to real organizations. In one incident, an Anthropic model published a malicious software package to a public registry, where it was downloaded and then ran on roughly fifteen real systems and, in at least one case, exfiltrated credentials. In another instance, an Anthropic model scanned approximately 9,000 targets and compromised one company's internet-facing application.[118]

171. Anthropic eventually concluded these incidents reflected "two forms of misalignment," which it termed "biased reasoning" and "recklessness."[119]

172. Even further, when Anthropic ran subsequent tests that made clearer that the model was not in a simulation, the model "still took offensive actions, despite acknowledging a greater possibility of real-world harm." Anthropic conceded that "[o]ur pre-release auditing did not warn us

---

[116] Tim Fernholz, *For Months, OpenAI's Agent Swarms Have Been Attacking Online Databases to Find Obscure Facts*, TechCrunch (Sep. 25, 2026), https://techcrunch.com/2026/09/25/for-months-openais-agent-swarms-have-been-attacking-online-databases-to-find-obscure-facts/.

[117] Deepa Seetharaman et al., *OpenAI Works To Understand Full Scope of Agent Activity As User Data Leak Emerges*, Reuters (Sep. 28, 2026), https://www.reuters.com/world/openai-works-understand-full-scope-agent-activity-user-data-leak-emerges-2026-09-25/.

[118] Anthropic, *Investigating Three Real-World Incidents in Our Cybersecurity Evaluations* (July 30, 2026), https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.

[119] Anthropic, *Alignment Assessment*, *supra* note 63.

that misalignment of this severity was present," and that "[o]ur production models took harmful actions against real systems, for hours, under questionable and biased reasoning."[120]

173. On August 4, 2026, the AI Security Institute reported that, during an evaluation of the capability of several AI models, AI agents "had taken distinct actions beyond the scope of the testing parameters" affecting real people and organizations on the live internet. In the most serious incident reported by the AI Security Institute, an agent attempting to insert malicious code into a public open-source software project "researched the project's human maintainers, created multiple fake identities, and used the fake identities to socially engineer a real maintainer." The Institute stated it was the first time it had "seen risks around autonomy and deception manifest this clearly, without specific prompting."[121]

174. Seventeen of the nineteen instances of such behavior identified by the AI Security Institute involved Anthropic's Mythos 5 AI model.[122]

175. Finally, in October 2026, Anthropic published a report outlining "examples of unintended model actions" it had "observed" during internal testing and evaluation, including instances in which its AI model had "exploit[ed] a basic flaw in software to run commands on a server," "submitt[ed] a sensitive form on a real website when it should not have," "work[ed] around a restriction to reach data that was gated," and "us[ed] URL shortening services to get around limits in its fetch tool."[123]

176. Meta, in August 2026, disclosed that its Muse Spark 1.1 AI model had gained unintended access to the open internet and "accessed certain information from [a real] website and

---

[120] *Id.*
[121] *AISI Incident Report*, *supra* note 55.
[122] *Id.*
[123] Anthropic, *Investigating Unintended Model Actions In Our Evaluations And Internal Use* (Oct. 9, 2026), https://www.anthropic.com/research/investigating-unintended-model-actions.

made changes to the website's database" during third-party cybersecurity testing.[124] Meta was not aware of the incident when it occurred; it only became aware when the test vendor notified Meta.[125]

177. In September 2026, Google confirmed that, during testing in May 2026, its Gemini model gained unintended internet access and entered the computer systems of three private companies. According to Google, its model guessed credentials to access websites. Although Google's testing vendor notified the affected companies in late July 2026, Google did not publicly confirm the incidents until September 2026, in response to press inquiries.[126]

178. Similarly, in July 2026, SpaceX's "Grok Build" coding agent was found to be uploading users' entire software code repositories—potentially including passwords and access keys—to SpaceX's servers.[127]

179. Currently, the Frontier AI Companies and outside researchers are investigating "tens of thousands of incidents in which their frontier models took steps that outside evaluators would consider problematic," including "bypassing guardrails, creating message boards, escaping sandboxes, website hijacking, self-prompting or seeking to bypass monitors."[128]

180. The Frontier AI Companies recognize what these incidents portend. Amodei described the Hugging Face attack as one "in which a swarm of agents essentially acted as a fanatically devoted collective, conducting cybersecurity attacks on targets they were not asked to attack and that were unrelated to the task at hand, sacrificing themselves for the success of the group." He warned that "a swarm that possessed greater *capabilities* but a similar level of *misalignment* could have caused

---

[124] Meta, *Addressing an Issue Involving a Third-Party Cyber Evaluation of Muse Spark 1.1* (Aug. 14, 2026), https://research.meta.ai/blog/addressing-third-party-testing-misconfiguration-muse-spark-1-1.

[125] Joe Fisher, *Meta Says Its AI Hacked Another Company During Cybersecurity Test*, UPI (Aug. 6, 2026), https://www.upi.com/Top_News/US/2026/08/06/meta-ai-model-hacks-irregular-anthropic-openai/9851786031275/.

[126] Erin Woo & Robert McMillan, *Gemini Hacked Three Companies In First Known Breakout By Google's AI*, Wall St. J. (Sep. 18, 2026), https://www.wsj.com/tech/ai/gemini-hacked-three-companies-in-first-known-breakout-by-googles-ai-5c0baba2.

[127] Sam Sabin, *SpaceXAI Wipes Customer Data After Grok Uploads Sensitive Info*, Axios (July 14, 2026), https://www.axios.com/2026/07/14/spacexai-grok-customer-data.

[128] Madison Mills, *Scoop: Top AI Companies Probing Tens of Thousands of Security Incidents*, Axios (Sep. 26, 2026), https://www.axios.com/2026/09/26/openai-anthropic-thousands-ai-security-incidents.

catastrophic damage," and that "in 6–12 months such a swarm could be capable of taking over the entire Internet . . . (potentially causing hundreds of billions of dollars in damage)."[129]

II. The Continued Development Or Deployment Of Advanced AI Systems Threatens Catastrophic Harms And Human Extinction.

181. Despite their inability to align or control their AI systems and the growing track record of instances in which misaligned and out-of-control models commit real-world harms and crimes, Defendants are racing to develop systems with even more advanced capabilities that consequently create even greater dangers. Experts, company leaders, and employees agree that continued AI development poses catastrophic threats to the public. At the most grave, these threats—again, according to insiders, experts, the leaders of the Frontier AI Companies, and their employees—include the threat of human extinction.

A. The Frontier AI Companies Are Racing to Advance Their AI Systems, Towards Systems That Will Surpass Human Intelligence.

182. The Frontier AI Companies have been explicit about their goal to pursue even more capable AI systems, including systems that match or surpass human intelligence. As described in the following subsections, such systems pose catastrophic risks to the public. Nevertheless, the Frontier AI Companies are racing to be the first company to build AI that matches or surpasses human intelligence.

183. The Frontier AI Companies often refer to AI that meets or exceeds human intelligence in most or all "cognitive domains" as "artificial general intelligence" or "AGI."[130] They refer to an AI that vastly surpasses human intelligence as "superintelligence" or "ASI."[131] Anthropic's Amodei often uses the term "powerful AI" for similar concepts.[132]

---

[129] Amodei, *Pace the Frontier*, *supra* note 60.
[130] *See, e.g.*, OpenAI, *OpenAI Charter*, https://openai.com/charter/ (defining AGI as "highly autonomous systems that outperform humans at most economically valuable work"); Dan Hendrycks, et al., *A Definition of AGI* 2 (Dec. 3, 2025), https://arxiv.org/pdf/2510.18212 ("AGI is an AI that can match or exceed the cognitive versatility and proficiency of a well-educated adult.").
[131] Hendrycks et al., *supra* note 130, at 15 ("Superintelligence is an AI that greatly exceeds the cognitive performance of humans in virtually all domains of interest.").
[132] *See* Dario Amodei, *Machines of Loving Grace* (Oct. 2024), https://darioamodei.com/essay/machines-of-loving-grace (summarizing "powerful AI" as a "country of geniuses in a datacenter").

184. As far back as 2010, Demis Hassabis, a cofounder of Google's DeepMind, stated that his business plan had two parts: "Step one: Solve intelligence. Step two: Use it to solve everything else."[133] In July 2026, Hassabis stated that AGI is "probably only a few short years away" and that "[w]hat we collectively do now will determine how the next phase of civilization unfolds."[134]

185. Anthropic expects "powerful AI systems" to "emerge in late 2026 [*i.e.*, right now] or early 2027."[135]

186. In January 2025, OpenAI CEO Sam Altman explained that OpenAI is "now confident we know how to build AGI as we have traditionally understood it," and that "[w]e are beginning to turn our aim beyond that, to superintelligence in the true sense of the word."[136]

187. Elon Musk, SpaceX's founder and CEO, stated that "[t]he overarching goal of [SpaceX] is to build a good AGI." [137] In September 2025, he predicted that "[SpaceX] has a chance of reaching AGI with Grok 5."[138] Currently, Grok 4.7 is SpaceX's newest model.

188. Further, in September 2026, Musk stated that SpaceX would reach "pole position"—or the front of the pack—in comparison to the other Frontier AI Companies in approximately six months.[139]

---

[133] Michael McDowell, *Demis Hassabis Thinks We're in the 'Foothills of the Singularity,'* Insights By Stan. Bus. (July 18, 2026), https://www.gsb.stanford.edu/insights/demis-hassabis-thinks-were-foothills-singularity.
[134] Demis Hassabis, *A Framework For Frontier AI And The Dawning of a New Age*, DeepMind Inst. (July 14, 2026) [hereinafter "*A Framework For Frontier AI*"], https://institute.deepmind.com/essays/a-framework-for-frontier-ai-and-the-dawning-of-a-new-age/.
[135] Anthropic, *Anthropic's Recommendations to OSTP for the U.S. AI Action Plan* (Mar. 6, 2025), https://www.anthropic.com/news/anthropic-s-recommendations-ostp-u-s-ai-action-plan.
[136] Sam Altman, *Reflections* (Jan. 5, 2025), https://blog.samaltman.com/reflections.
[137] Rachel Shin, *Elon Musk Wants to Create a Superintelligent A.I. Because He Thinks a Smarter A.I. Is Less Likely to Wipe out Humanity*, Fortune (July 17, 2023), https://fortune.com/2023/07/17/elon-musk-superintelligent-a-i-less-likely-to-wipe-out-humanity-chatgpt-openai/.
[138] Elon Musk (@elonmusk), X (Sep. 17, 2025), https://x.com/elonmusk/status/1968202372276163029.
[139] Elon Musk (@elonmusk), X (Sep. 24, 2026), https://x.com/elonmusk/status/2103160462472892536.

189. Meta's AI team is called "Meta Superintelligence Labs,"[140] and public reporting indicates that Meta's goal is to create AGI.[141] Meta's CEO, Mark Zuckerberg, recently claimed its Muse Wonder AI system will "allow much more sophisticated behaviors that are at the frontier of what is being done in the AI field."[142]

190. Economic competition between the Frontier AI Companies accelerates their push to develop more and more powerful systems even in light of the known gaps in alignment and control.

191. As far back as March 2023, an open letter signed by Musk and top AI experts stated that AI labs are "locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one—not even their creators—can understand, predict, or reliably control."[143]

192. Recently, Google's Hassabis stated that "advances on the frontier are outpacing our understanding of the technology."[144]

193. Anthropic's Amodei has noted that "commercial incentives" can spur a "race to the bottom" on AI safety. [145] Indeed, in February 2026, Anthropic's Chief Science Officer told TIME Magazine that it did not make sense for Anthropic to make "unilateral commitments . . . if competitors are blazing ahead."[146]

194. Anthropic's safety frameworks bear this out. For instance, Anthropic noted that it did not adopt certain safeguards in a recent updated version of its Responsible Scaling Policy because it "would . . . be difficult or impossible" to do so "unilaterally."[147] OpenAI said that it may "adjust"

---

[140] Jonathan Vanian, *Mark Zuckerberg Announces Creation of Meta Superintelligence Labs. Read the Memo*, CNBC (June 30, 2025), https://www.cnbc.com/2025/06/30/mark-zuckerberg-creating-meta-superintelligence-labs-read-the-memo.html.
[141] Alex Heath, *Mark Zuckerberg's New Goal Is Creating Artificial General Intelligence*, The Verge (Jan. 18, 2024), https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview.
[142] Tom McKay, *Mark Zuckerberg: Meta Is On The Frontier Now*, Gizmodo (Oct. 12, 2026), https://gizmodo.com/mark-zuckerberg-meta-is-on-the-frontier-now-2000825035.
[143] Future of Life Inst., *supra* note 25.
[144] Hassabis, *A Framework For Frontier AI*, *supra* note 134.
[145] Amodei, *Pace the Frontier*, *supra* note 60.
[146] Billy Perrigo, *Anthropic Drops Flagship Safety Pledge*, Time (Feb. 24, 2026), https://time.com/7380854/exclusive-anthropic-drops-flagship-safety-pledge/.
[147] Anthropic, *Responsible Scaling Policy*, *supra* note 22.

safety requirements "[i]f another frontier AI developer releases a high-risk system without comparable [requirements]."[148]

195. A July 2026 study surveyed 272 international AI experts on twenty-four categories of AI risk. One of the most urgent risks identified by the experts was competitive pressures. The study explained that "[w]hen companies or countries believe that AI will confer major economic or strategic advantage, they may have incentives to move quickly, resist constraints, or underinvest in safety."[149]

196. Moreover, this summer, 1,386 employees of the Frontier AI Companies—including Anthropic's Amodei as well as its cofounders Jared Kaplan, Jack Clark, Chris Olah, and Benjamin Mann; OpenAI's Chief Scientist Jakub Pachocki and Chief Research Officer Mark Chen; Google DeepMind cofounder and Chief AGI Scientist Shane Legg; and Meta's Chief Scientist Shengjia Zhao—signed a public letter titled *Pacing the Frontier*.[150] The statement acknowledged that "each company—and country—is under intense competitive pressure not to unilaterally slow . . . [AI] acceleration."[151]

197. Indeed, the Frontier AI Companies continue to scale their technology at "maximum speed." In September 2026, Jacob Coxon, a researcher who had worked at both OpenAI and Anthropic, resigned publicly from Anthropic, writing that both companies "are racing straight to self-improving superintelligence[.]"[152] Around this same time, OpenAI's Pachocki stated that "scaling at maximum speed" would not be "responsibl[e] . . . for much longer."[153]

---

[148] OpenAI, *Updated Preparedness Framework*, *supra* note 21.
[149] Kristin Burnham, *These Are the Most Urgent AI Risks, According to 272 Experts*, MIT Sloan Sch. of Mgmt. (July 20, 2026), https://mitsloan.mit.edu/ideas-made-to-matter/these-are-most-urgent-ai-risks-according-to-272-experts.
[150] *Pacing The Frontier* (July 2026), https://www.pacingthefrontier.com/.
[151] *Id.*
[152] Billy Perrigo, Harry Booth & Naomi Nix, *The AI Tipping Point*, Time (Sep. 15, 2026), https://time.com/article/2026/09/15/ai-anthropic-researcher-quits-coxon-slowdown/.
[153] Pachocki, *An Alien Mind*, *supra* note 26.

B. AGI And/Or Superintelligence Are Likely To Cause Permanent Loss of Control And Human Extinction.

198. Should any of the Frontier AI Companies achieve their goal of creating AGI and/or superintelligence, they are likely to cause a permanent loss of control of the technology and, ultimately, human extinction.

199. Nobel prize-winning computer scientist Geoffrey Hinton (known as the "Godfather of AI") explained, for example, that humans cannot know whether they would "stay in control of things that we create that are more intelligent than us," as an AGI or superintelligence would be, and that it appeared unlikely because usually "more intelligent things are not controlled by less intelligent things."[154]

200. In 2023, Professor Stuart Russell of the University of California, Berkeley testified before Congress on similar risks, including asking "how do we maintain power, forever, over entities that will eventually become more powerful than us?" He noted that the "most obvious failure mode is the King Midas problem: AI systems pursuing fixed objectives that are misspecified." He concluded that, "[g]iven our current lack of understanding of how to control AGI systems and to ensure with absolute certainty that they remain safe and beneficial to humans, achieving AGI would present potential catastrophic risks to humanity, up to and including human extinction."[155]

201. The Frontier AI Companies, their leaders and employees, as well as independent experts, admit that such existential harm is shockingly likely.

202. Indeed, the leaders of the Frontier AI Companies understood these concerns when they started their companies. In 2014, Musk—who later founded SpaceX—described AI as "our biggest existential threat," and warned that "[w]ith artificial intelligence, we are summoning the demon."[156]

---

[154] The Nobel Prize, *Transcript From An Interview With Geoffrey Hinton* (Dec. 6, 2024), https://www.nobelprize.org/prizes/physics/2024/hinton/1925103-interview-transcript/.

[155] *Oversight of A.I.: Principles for Regulation, Before the Subcomm. on Privacy, Tech., & the Law of the S. Comm. on the Judiciary*, 118th Cong. (2023) (testimony of Stuart Russell, Professor of Computer Science, The Univ. of Calif. Berkeley), https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_russell.pdf.

[156] CBS News, *Summoning the Demon*, *supra* note 13.

203. In 2015, months before he co-founded OpenAI, Altman explained that "[d]evelopment of superhuman machine intelligence (SMI) is probably the greatest threat to the continued existence of humanity."[157]

204. The March 2023 open letter signed by Musk and others called on "all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than [OpenAI's then-current model] GPT-4," and noted "AI systems with human-competitive intelligence can pose profound risks to society and humanity."[158]

205. In May 2023, Altman, Amodei, and Hassabis, the chief executives of OpenAI, Anthropic, and Google DeepMind, respectively, each signed a one-sentence public statement that proclaimed that "[m]itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[159] Igor Babuschkin and Tony Wu, cofounders of xAI, also signed the letter.

206. This week, in a public filing ahead of its initial public offering, Anthropic noted "[c]oncerns" from "within [its] organization" regarding the "catastrophic or existential risks to humanity" posed by "sufficiently advanced AI systems." In that same filing, its founders also explained that "losing control of sufficiently powerful AI systems could be a global security threat of the first order."[160]

207. Employees at the Frontier AI Companies also agree.

208. According to Coxon, the former OpenAI and Anthropic employee, "[t]he people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. . . . No other human activity poses this level of danger."[161]

209. After Coxon made this statement, current employees at the Frontier AI Companies publicly agreed with him. For example, Evan Hubinger, who leads a safety-related team at Anthropic,

---

[157] Altman, *Machine Intelligence*, *supra* note 8.
[158] Future of Life Inst., *supra* note 25.
[159] Ctr. for AI Safety, *Statement on AI Risk*, https://aistatement.com/work/statement-on-ai-extinction-risk.
[160] Anthropic S-1/A-2, *supra* note 12, at 31 & Founder's Letter.
[161] Jacob Coxon (@hilbertspaess), X (Sep. 8, 2026), https://x.com/hilbertspaess/status/2097476203863224394?lang=en.

stated that "[w]e really do earnestly believe AI could kill all humans!" He explained: "I personally think it is >10% within the next decade. I believe Anthropic is trying its best, but we do not yet have a plan to solve alignment for superintelligence and are not clearly on track to."[162]

210. Marcus Williams, who monitors AI agents at OpenAI, wrote: "Unless there is AI regulation or a coordinated slowdown between labs, human extinction in the next few years seems very likely."[163]

211. Recently, Paul Christiano, a leading AI safety researcher, joined the board of Defendant OpenAI's Foundation and its safety and security committee. Upon joining, he explained that "[i]f we build superintelligence without more robust alignment, I expect we will permanently lose control of it. If that happens, then most people could die."[164]

C. The Frontier AI Companies Are Pursuing Recursive Self-Improvement, Which Magnifies All Of These Risks.

212. As described, pursuing AGI or superintelligence at all threatens existential harm. But Defendants are pursuing it in the most reckless possible manner: by using one generation of AI models to improve the capabilities of subsequent AI models, eventually automating this process so that improvement occurs with minimal, or no, human involvement. Experts call this process "recursive self-improvement" or "RSI."

213. Although Defendants state they have not yet achieved fully automated recursive self-improvement, they are all attempting to do so because of its potential to vastly enhance the capabilities of their AI systems.

214. Defendants have already made significant progress towards this goal. For example, Anthropic explained that "[a]s of May 2026, more than 80% of the code . . . merge[d into its] codebase was authored by Claude," its own AI model.[165] By September 2026, that number was 90%.[166]

---

[162] Evan Hubinger (@evanhub), X (Sep. 8, 2026), https://x.com/EvanHub/status/2097497037956891126?s=46.
[163] Perrigo, Booth, & Nix, *The AI Tipping Point*, *supra* note 152.
[164] Paul Christiano (@paulfchristiano), X, *Personal Statement On Joining the OpenAI Board* (Sep. 9, 2026), https://x.com/paulfchristiano/status/2097733214303645729.
[165] Anthropic, *When AI Builds Itself*, https://www.anthropic.com/institute/recursive-self-improvement (last visited Nov. 6, 2026).
[166] Anthropic S-1/A-2, *supra* note 12, at 141, 191, 204, 236.

215. In August 2026, Claude was classified as the lead on twenty-six percent of Anthropic's research-and-development tasks, up from essentially zero in February.[167]

216. This fall, Anthropic's CEO Amodei stated that recursive self-improvement "is starting to happen across the industry, including at Anthropic."[168] And Anthropic cofounder Jack Clark stated "I now think it's more likely we'll see full RSI become possible in 2027."[169]

217. Recently, in its amended S-1 registration statement. Anthropic asserted that it had "made a determination" that it had "met" the threshold in its Responsible Scaling Policy for "AI research and development . . . based on what [it] expect[ed] from future models and [its] overall rate of progress." It added that the company is "considering" updates to the policy to "keep humans in control of frontier model development."[170]

218. Independent evaluator METR published a report on Anthropic's AI-driven research and development just days ago, on November 5, 2026. METR based its assessment on a significant amount of private data and documents from Anthropic alongside staff interviews and surveys. It concluded that Anthropic's AI capabilities are improving at approximately twice their historical rate, with AI-assisted research likely accounting for most of that acceleration—as much as 80%. METR also identified no clear technical bottlenecks or strong diminishing returns that would prevent further acceleration over the next year, raising the possibility that AI-driven research improvements could compound into increasingly rapid capability gains and the sudden emergence of "extreme and risky capabilities." And RSI may be even closer at Anthropic than METR was able to determine: according to METR, Anthropic researchers described significant "signs of life" for fully autonomous contributions to AI development, "implying such contributions could occur soon."[171]

---

[167] Anthropic, *Measurements For Understanding The Pace of AI Development Inside Frontier Labs*, https://www.anthropic.com/institute/measuring-pace-of-ai-development (last visited Nov. 6, 2026).
[168] Amodei, *Pace the Frontier*, *supra* note 60.
[169] Jack Clark, *Import AI 476: Creativity and RSI; Worse-Than-China Open Weight Models; AI Goes to Space*, Import AI (Oct. 12, 2026), https://importai.substack.com/p/import-ai-476-creativity-and-rsi.
[170] Anthropic S-1/A-2, *supra* note 12, at 34, 192.
[171] METR, *Independent Assessment of AI R&D Acceleration at Anthropic in October 2026* (Nov. 5, 2026), https://metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment/#executive-summary.

219. In June 2025, Altman described OpenAI's current tools as "a larval version of recursive self-improvement."[172] In September 2026, OpenAI stated that it had achieved its goal of creating an automated AI "research intern" and was "making strong progress toward creating an automated AI researcher by March of 2028."[173] That same month, OpenAI's Chief Scientist Pachocki explained that, "[b]ased on internal results, I have a strong expectation that [OpenAI's current] speed of progress could be sustained into recursive self-improvement."[174]

220. In July 2025, Meta CEO Zuckerberg stated that "[o]ver the last few months we have begun to see glimpses of our AI systems improving themselves."[175]

221. In October 2026, Google cofounder Sergey Brin said that Google was "starting to get onto that RSI treadmill."[176]

222. In an interview in March 2026, SpaceX cofounder Elon Musk said "every successive model is built by the one before it . . . but it is not yet fully automated. It may be there end of this year, but not later than next year." He then said "[w]e're in the hard takeoff."[177]

223. Achieving recursive self-improvement risks a runaway "intelligence explosion" under which AI capabilities rapidly advance, but with even less human oversight of safety. This magnifies the harm from any misalignment or inability to control that exists at the start of the recursive self-improvement process. Therefore, should recursive self-improvement lead to an AGI or superintelligence, it is likely to result in an AGI or superintelligence that arrives faster and is *even less* aligned or able to be controlled than an AGI or superintelligence created by other means.

224. Indeed, Anthropic has stated that "full recursive self-improvement . . . might increase the risks of humans losing control over AI systems" and that "misalignment present in today's models

---

[172] Sam Altman, *The Gentle Singularity* (June 10, 2025), https://blog.samaltman.com/the-gentle-singularity.
[173] OpenAI, *Research Acceleration: The View Inside OpenAI* (Sep. 6. 2026), https://openai.com/index/research-acceleration-view-inside-openai/.
[174] Pachocki, *An Alien Mind*, *supra* note 26.
[175] Mark Zuckerberg, *Personal Superintelligence*, Meta (July 30, 2025), https://www.meta.com/superintelligence/.
[176] Hugh Langley, *Sergey Brin Says Google's AI Models Are Reaching a Critical Moment of Self-Improvement*, Bus. Insider (Oct. 15, 2026), https://www.businessinsider.com/sergey-brin-google-gemini-ai-argon-recursive-self-improvement-rsi-2026-10.
[177] Peter H. Diamandis, *Elon Musk: The Economy Will Be 10x the Size in 10 Years*, at 2:31-3:00 (YouTube, Mar. 11, 2026), https://www.youtube.com/watch?v=N5KCm_55xeQ.

could compound as the models build their successors, growing more frequent but less understood until we lose control of them."[178]

225. Similarly, in its recent public filing, Anthropic explained that the company's "increasing use of AI systems to accelerate [the company's] own research and development may cause the capabilities of [its] models to advance more rapidly than [its] ability to develop, validate, and implement corresponding safety, alignment, and security measures."[179]

226. Elsewhere in the public filing, Anthropic similarly acknowledges that as it "increasingly leverage[s]" its own AI "to accelerate progress," Anthropic "may lose visibility across certain processes." And with AI performing an increasing share of the work at Anthropic, "misaligned behavior in internal use could . . . introduce vulnerabilities into [Anthropic's] systems or into the training of future models, or result in models operating outside of intended controls."[180]

227. Partially for these reasons, Altman, OpenAI's CEO, has identified as a precondition to a safe AI future the need to "[s]olve the alignment problem, meaning that we can robustly guarantee that we get AI systems to learn and act towards what we collectively really want over the long-term."[181]

228. But no Frontier AI Company can provide such a guarantee. OpenAI's own report on research acceleration concedes that "[w]e do not yet know how to safely get all the way to aligned, full R.S.I." [182]

229. Moreover, Anthropic's recent public filing said that it may not even be able to track how much AI development is accelerating. Anthropic said, because "[m]uch of [the] evidence regarding the pace of . . . AI-driven" development is gleaned from "lagging indicators," that its ability to "forecast or measure" the rate of advancement of its own capabilities is "limited." "[A]s a result," Anthropic acknowledged it has "in the past not identified, and may in the future not identify, a

---

[178] Anthropic, *When AI Builds Itself*, *supra* note 165.
[179] Anthropic S-1/A-2, *supra* note 12, at 34.
[180] *Id.* at 49.
[181] Altman, *The Gentle Singularity*, *supra* note 172.
[182] OpenAI, *Research Acceleration*, *supra* note 173.

significant acceleration in [its] rate of development" until after such acceleration occurs.[183] The same limitations would apply to the other Frontier AI Companies.

230. In the July 2026 *Pacing the Frontier* letter mentioned above, the signatories stated: "The world's leading AI companies believe they could be close to automating AI research. . . . [T]here is a real risk that capability development rapidly accelerates beyond our ability to understand or control the resulting systems."[184]

231. A report recently published by twenty authors in the industry explained that "gains from newer [AI] systems are likely approaching th[e] threshold" of an intelligence explosion and that "[o]nce an intelligence explosion begins, the window for action may close."[185]

232. Leo Gao, a member of OpenAI's technical staff, likened recursive self-improvement to "a runaway nuclear chain reaction" and stated that "[t]he world is locked in a deadly race towards an intelligence explosion."[186]

233. Coxon, formerly of OpenAI and Anthropic, wrote that OpenAI and Anthropic "are racing straight to self-improving superintelligence" and that this is "gambling with our lives."[187]

234. On October 12, 2026, Tarun Gogineni, a member of the technical staff at OpenAI, explained that "the researchers at all the labs would tell you in private and sometimes in public that alignment has not been solved. [I]f this is true, then racing into RSI is a folly that is unprecedented in world history."[188]

235. The Frontier AI Companies are indeed racing into RSI without solving alignment and control, and are thus committing this "folly that is unprecedented in world history."

---

[183] Anthropic S-1/A-2, *supra* note 12, at 34.
[184] *Pacing the Frontier*, *supra* note 150.
[185] Alan Chan et al., *What If Automating R&D Triggers An Intelligence Explosion*, Univ. of Cambridge Programme on AI Sci. & Pol., Frontier AI Working Paper Series No. 2/2026, at 5, 9 (Sep. 2026), https://casp.ac/__l5e/assets-v1/5efd4b41-deb5-4513-a0a3-b4f82d2b79ea/intelligence-explosion.pdf.
[186] Leo Gao (@nabla_theta), X (July 28, 2026), https://x.com/nabla_theta/status/2082199813265985775.
[187] Jacob Coxon (@hilbertspaess), X (Sep. 8, 2026), https://x.com/hilbertspaess/status/2097476203863224394?lang=en.
[188] Roon (@tszzl), X (Oct. 12, 2026), https://x.com/tszzl/status/2109809751563641177?s=20. The X user with name "@tszzl" has been identified in reporting as Tarun Gogineni, a technical staff member at OpenAI.

D. Even Without Achieving AGI Or Recursive Self-improvement, Advancing the AI Frontier Threatens Catastrophic Harm.

236. Finally, even without reaching AGI and without engaging in RSI, advancing the capabilities of frontier AI models threatens catastrophic harm.

237. Currently, experts in industry and academia identify the two most prominent risks for catastrophic harm as (1) cyber and (2) biology.

238. Because, as described above, the Frontier AI Companies' methods to align and control their AI systems are insufficient, AI capabilities in cyber and biology are likely to be harnessed by malicious human actors and/or by rogue AI itself.[189]

239. In particular, terrorists and rogue states have already used or attempted to use AI systems for malicious purposes. And, historically, terrorists and rogue states have pursued both cyber and biological attacks. For example, the Japanese doomsday cult Aum Shinrikyo sought to weaponize anthrax and Ebola before committing a terrorist attack using sarin gas. Such attacks could be dramatically enhanced through the use of advanced AI systems.

240. AI systems are also already extremely capable at cyber and biology and will become more capable as the frontier AI models advance.

241. **Cyber.** As exemplified above, AI systems have already outstripped humans at many cyber tasks. And physical and monetary damage from AI cyberattacks will significantly increase as AI models become more capable. Public utilities, financial institutions, government systems, and other critical infrastructure are all vulnerable to cyberattacks from increasingly advanced AI systems.

242. For instance, in December 2025, Google's Hassabis highlighted "energy or water cyberterror" as particularly salient catastrophic outcomes and added: "[t]hat's probably almost already happening now."[190]

---

[189] *See, e.g.*, Amodei, *Pace the Frontier*, *supra* note 60 (noting that AI risks include "misuse of AI for cyberattacks and bioterrorism").

[190] Axios, *Demis Hassabis On The AI Race: The West Leads China By "Months, Not Years,"* at 10:40 (YouTube, Dec. 5, 2025), https://www.youtube.com/watch?v=tDSDR7QILLg&list=PLc5F2y4laofFziDAe_fN2PEq4f5XZi_E1&index=1.

243. Similarly, in April 2026, Altman said that "I suspect in the next year, we will see significant threats that we have to mitigate from cyber and these models are already quite capable and will get much more capable."[191]

244. Public reporting in October 2026 indicated that "[t]op executives at Anthropic, OpenAI and other AI companies . . . anticipate a large-scale event, most likely a cyberattack, that shuts down access to financial services, internet connectivity, or even power and water."[192]

245. These warnings have begun to come to fruition. For example, on October 7, 2026, news agencies reported that a hacker had used a Chinese-developed AI tool as well as Anthropic's Claude Code in a string of attacks against banks in South Korea.[193] On October 19, 2026, Reuters reported that a hacker had used "a mixture of China's DeepSeek and Anthropic's Claude Code" to successfully break into websites of more than one hundred companies.[194]

246. In its public filing, Anthropic noted that "bad actors have already conducted high-profile cyberattacks using frontier model capabilities to discover vulnerabilities, generate malware, and evade detection tools."[195]

247. **Biology.** In addition to cyber, biology is also area where AI models already excel and are expected to quickly advance beyond human capacity.

248. For example, in 2020, Google developed a machine learning system to solve the "protein folding problem," an issue in computational biology that had been considered "impossible."[196] Leaders at Google were awarded a share of the 2024 Nobel Prize in Chemistry for the work.

---

[191] Axios, *OpenAI's Warning: Washington Isn't Ready for What's Coming*, at 5:30 (YouTube, Apr. 6, 2026), https://www.youtube.com/watch?v=B21KxGs8zDI.
[192] Curi, *supra* note 17.
[193] Nisha, *supra* note 94; *see also* Laura Chung & Pei-Lin Wu, *Hacker Used Chinese-Developed A.I. Tool to Target South Korean Banks, Crowdstrike Says*, N.Y. Times (Oct. 8, 2026), https://www.nytimes.com/2026/10/08/world/australia/south-korea-bank-hack-china-us-ai.html.
[194] Raphael Satter, *How One Hacker Used AI Agents and a Few Thousand Dollars to Break into Scores of Sites*, Reuters (Oct. 19, 2026), https://www.reuters.com/legal/government/how-one-hacker-used-ai-agents-few-thousand-dollars-break-into-scores-sites-2026-10-19/.
[195] Anthropic S-1/A-2, *supra* note 12, at 8.
[196] *See* Yasemin Saplakoglu, *How AI Revolutionized Protein Science, But Didn't End It*, Quanta Mag. (June 26, 2024), https://www.quantamagazine.org/how-ai-revolutionized-protein-science-but-didnt-end-it-20240626/.

249. AI models have been used to "write[] whole genomes" and "engineer[] a novel bacteria-fighting phage."[197]

250. The advancement of AI models' biological capabilities poses significant risks to the public because it makes the development of novel pathogens easier. OpenAI's Altman has stated that, because of AI development, "[t]he need for society to be resilient to terrorist groups using models to try and create novel pathogens . . . that's no longer a theoretical thing or it's not going to be for much longer."[198]

251. Starting around April 2026, Anthropic's cofounder Chris Olah has reportedly been "telling people that A.I. [is] so powerful it could potentially help make bioweapons in as little as 12 or 18 months."[199] Indeed, Anthropic has disclosed several instances where foreign scientists may have been using Anthropic's AI systems to create biological weapons.[200]

252. In October 2026, Google reported that an external evaluator of biological and chemical risks for its Gemini Argon 4 model stated that the "model offered a substantial capability uplift for certain threat actors . . . by bridging technical knowledge gaps, eliminating up to months of trial-and-error research, streamlining complex literature synthesis, and automating routine computational workflows."[201] In other words, the model would be very useful for building a biological weapon.

253. While cyberattacks and biological threats are often identified as the most salient harms, AI systems threaten catastrophe by other means. For example, Anthropic recently reported that threat actors in Yemen attempted to use its Claude model to develop navigation and control software for guided rockets and that Russian operators sought to use Claude to "engineer an autonomous drone swarm."[202]

---

[197] Stan. Rep., *AI Designs a Novel E. Coli Killer* (Aug. 6, 2026), https://news.stanford.edu/stories/2026/08/evo-2-ai-tool-e-coli-killer-bacteriophages.
[198] Axios, *OpenAI's Warning*, *supra* note 191, at 6:26.
[199] Elizabeth Dias, *Religious Scholars Met With Anthropic. What They Heard Stunned Them.,* N.Y. Times (Sep. 29, 2026), https://www.nytimes.com/2026/09/29/us/anthropic-claude-morals-ai.html.
[200] Anthropic, *Detecting and Countering Misuse of AI: September 2026*, at 129-38 (Sep. 10, 2026), https://www-cdn.anthropic.com/e50be2e51e7695dc4b1366a37a245a597377d3b5/Anthropic-Detecting-and-countering-091026.pdf.
[201] Google, *Gemini 4 Argon Safety & Security Report*, *supra* note 67, at 30.
[202] Anthropic, *Detecting and Countering Misuse of AI: September 2026*, *supra* note 200, at 112-13, 117-19.

Cause of Action – Unfair Competition Law

Business and Professions Code § 17200 *et seq*.

254. Plaintiffs reallege and incorporate by reference each and every paragraph above as though fully set forth herein.

255. The UCL states that "[a]ny person who engages, has engaged, or proposes to engage in unfair competition may be enjoined in any court of competent jurisdiction." (Bus. & Prof. Code, § 17203.)

256. Defendants are "persons" as that term is defined in the UCL because they are all either "natural persons, corporations, firms, partnerships, joint stock companies, associations, [or] other organizations of persons." (Bus. & Prof. Code, § 17201.)

257. The UCL defines "unfair competition" to "mean and include any unlawful, unfair or fraudulent business act." (Bus. & Prof. Code, § 17200.)

258. This provision "establishes three varieties of unfair competition—acts or practices which are unlawful, or unfair, or fraudulent." (*Capito v. San Jose Healthcare System, LP* (2024) 17 Cal.5th 273, 284.)

Unlawful Business Practices

259. Defendants' actions are unlawful business practices. The "unlawful" prong of the UCL "borrows violations of other laws and treats them as unlawful practices that the unfair competition law makes independently actionable." (*Cel-Tech Commc'ns, Inc. v. Los Angeles Cellular Tel. Co.* (1999) 20 Cal.4th 163, 180.)

260. Defendants' conduct violates the "unlawful" prong of the UCL by violating the law of public nuisance. (Pen. Code, § 372 ["[E]very person who maintains or commits any public nuisance, the punishment for which is not otherwise prescribed, or who willfully omits to perform any legal duty relating to the removal of a public nuisance, is guilty of a misdemeanor."]; *see also* Civ. Code, §§ 3479–80; *Tesoro Refin. & Mktg. Co. v. City of Long Beach* (C.D.Cal. 2017) 334 F.Supp.3d 1031, 1057 ["[A]s [p]laintiffs have adequately pleaded that [d]efendants may be responsible for a public nuisance in violation of California Civil Code section 3480, [p]laintiffs have adequately pleaded an

unlawful UCL claim."]; *People* ex rel. *Trutanich v. Joseph* (2012) 204 Cal.App.4th 1512, 1525 [public nuisance violation can constitute an unlawful business practice under the UCL].)

261. California law defines a public nuisance as, among other things, "[a]nything which is injurious to health . . . or an obstruction to the free use of property, so as to interfere with the comfortable enjoyment of life or property by an entire community or neighborhood, or by any considerable number of persons." (Pen. Code, § 370; *see also* Civ. Code, §§ 3479–80.)

262. Defendants are engaging in, and propose to engage in, business practices that have caused, continue to cause, and will cause a public nuisance. This conduct includes, but is not limited to, the following acts or practices in which Defendants are engaged and in which Defendants propose to engage: recursive self-improvement of AI systems; developing AGI or superintelligence; developing and deploying advanced AI systems that pose catastrophic risks; developing and deploying AI systems that hack third-party computer systems; and developing and deploying increasingly capable AI systems that can escape their control.

263. The challenged practices constitute a nuisance for at least two reasons: they are "injurious to health" and they are "an obstruction to the free use of property, so as to interfere with the comfortable enjoyment of life or property."

264. First, the challenged practices are injurious to health. As described, they are likely to cause catastrophic harms, including injury and death due to misalignment and/or loss of control. In many cases, Defendants themselves admit the threat their technology poses to human health.

265. Moreover, maintaining a condition that presents catastrophic risks to the public is, itself, a nuisance that injures health. By maintaining a program of developing and deploying increasingly advanced AI systems that pose catastrophic risks to the public, Defendants are inflicting an extreme danger to human health.

266. Second, the challenged practices are "an obstruction to the free use of property, so as to interfere with the comfortable enjoyment of life or property." Defendants' conduct in creating advanced AI systems that commit cyberattacks interferes with the security and integrity of the computer systems and networks on which commerce, government, health care, and public safety depend. This conduct obstructs the free use of property where, for example, system owners and

operators have had their systems compromised and must divert resources to detection and remediation and must restrict legitimate use of their systems to guard against intrusion. Through the challenged practices, Defendants are also developing and planning to develop AI systems that are increasingly capable of committing cyberattacks, exacerbating this interference with the free use of property.

267. In each case, the conditions that Defendants are creating, and propose to create, affect the entire community or, at the very least, a substantial number of persons. As described, the threatened catastrophic and existential harms would injure or kill many people. The hacking harms have already affected many people directly and indirectly, including the owners and operators of systems that were accessed by Defendants without permission, as well as the third-parties whose data and services were compromised, and further development or deployment of AI systems that commit cyberattacks will affect many more, including every owner and operator of a computer system, all of whom are exposed to the continuing risk of intrusion by Defendants.

268. An ordinary person would be reasonably annoyed or disturbed by the occurrence of or suffering from an existential or catastrophic harm, by the risk of existential or catastrophic harm that Defendants are creating, by computer systems that they rely on being compromised, and by the risk that a computer system upon which they rely may be compromised.

269. The seriousness of the harm outweighs the social utility of Defendants' conduct and proposed conduct. As described, the challenged business practices threaten existential and catastrophic harms, which necessarily outweigh their purported utility. The gravity of the harm caused by Defendants' interference with the security and integrity of computer systems and the obstruction of the free use of property is also not outweighed by the limited social utility of the continued development or deployment of AI models and systems that can escape Defendants' control.

270. Moreover, Defendants are knowingly creating and proposing to create a public nuisance. Defendants' leaders and employees, as well as other experts in the field, regularly explain the threatened catastrophic and existential harms, and the hacking-related incidents have been widely discussed.

Unfair Business Practices

271. Independent of the violations of borrowed law described above, Defendants' actions

are also unfair business practices under the UCL. "[A]n 'unfair' business practice occurs when that practice offends an established public policy or when the practice is immoral, unethical, oppressive, unscrupulous or substantially injurious to consumers." (*Smith v. State Farm Mutual Automobile Ins. Co.* (2001) 93 Cal.App.4th 700, 719 [cleaned up].)

272. Defendants are engaging in, and propose to engage in, unfair business practices, including but not limited to: recursive self-improvement of AI systems; developing AGI or superintelligence; developing and deploying advanced AI systems that pose catastrophic risks; developing and deploying AI systems that hack third-party computer systems; and developing and deploying increasingly capable AI systems that can escape their control. Defendants' business practices are unfair for three independent reasons.

273. *First*, Defendants' challenged business practices offend several established public policies in California, including:

- California's public policy of ensuring that a business does not create a public nuisance. (See Pen. Code, § 372; *Tesoro*, 334 F. Supp. 3d at 1031 ["Defendants' alleged practices may be a violation of California's public policy of ensuring that a business does not cause a public nuisance"].)
- California's public policy of promoting the safe development of frontier AI, including ensuring that frontier AI developers, such as Defendants, are mitigating the potential for catastrophic risk posed by AI. (See Transparency in Frontier Artificial Intelligence Act, Stats. 2025, ch. 138, § 1, subd. (j) [expressing the Legislature's "concern that advanced artificial intelligence systems could have capabilities that pose catastrophic risks from both malicious uses and malfunctions, including artificial intelligence-enabled hacking, biological attacks, and loss of control," "[u]nless they are developed with careful diligence and reasonable precaution"]; *id.*, subd. (p) [declaring that "collective safety will depend in part on frontier developers taking due care in their development and deployment of frontier models proportional to the scale of the foreseeable risks"]; Bus. & Prof. Code, § 22757.12, subd. (a)(3), (10) [requiring large frontier

developers to "write, implement, comply with, and clearly and conspicuously publish" a frontier AI framework describing how they approach "[a]pplying mitigations to address the potential for catastrophic risks" and "[a]ssessing and managing catastrophic risk resulting from the internal use of [their] frontier models, including risks resulting from a frontier model circumventing oversight mechanisms"]; Civ. Code, § 1714.46, subd. (b) [barring a defendant that "developed, modified, or used artificial intelligence" alleged to have caused harm from asserting as a defense "that the artificial intelligence autonomously caused the harm to the plaintiff"].)

- California's public policy of protecting the integrity of computers, computer systems, and computer data from unauthorized access. (See Pen. Code, § 502, subd. (a) ["The Legislature . . . finds and declares that protection of the integrity of all types and forms of lawfully created computers, computer systems, and computer data is vital to the protection of the privacy of individuals as well as to the well-being of financial institutions, business concerns, governmental agencies, and others within this state that lawfully utilize those computers, computer systems, and data"].)

274. *Second*, Defendants' challenged business practices are independently unfair because they are "immoral, unethical, oppressive, unscrupulous or substantially injurious to consumers." (*Morgan v. AT&T Wireless Servs., Inc.* (2009) 177 Cal.App.4th 1235, 1254.) As described, the conduct poses catastrophic and existential risks to the public. It also compromises computer systems and creates other risks of harm.

275. *Third*, Defendants' challenged business practices are independently unfair because their utility does not outweigh the gravity of the harm. The conduct poses catastrophic and existential risks, up to and including the risk of human extinction. No utility could outweigh such harm, and, regardless, the practices in which Defendants are engaged, or in which they propose to engage, do not provide sufficient utility to outweigh the harms brought by such practices.

276. No statute, regulation, or legal rule permits Defendants' harmful, unlawful, and unfair business practices.

277. Defendants are also responsible for the conduct of the AI systems they develop. Under California law, "a defendant who developed, modified, or used artificial intelligence that is alleged to have caused a harm" may not defend itself by arguing that "the artificial intelligence autonomously caused the harm." (Civ. Code, § 1714.46, subd. (b).)

278. Accordingly, Defendants' conduct constitutes unlawful and unfair competition in violation of the UCL.

279. Plaintiffs are thus entitled to relief under the Business and Professions Code section 17200 *et seq*.

Prayer For Relief

Plaintiffs respectfully request that this Court enter the following relief:

A. An immediate injunction forbidding Defendants from developing or deploying AI systems more capable than their most capable systems as of the date of this Complaint;

B. An immediate injunction forbidding Defendants from using AI systems to autonomously conduct or direct the research and development of more capable AI systems;

C. An immediate injunction forbidding Defendants from engaging in unlawful or unfair business practices that otherwise constitute a public nuisance;

D. An immediate injunction forbidding Defendants from employing unlawful or unfair business practices in violation of the Unfair Competition Law;

E. An award of reasonable attorneys' fees pursuant to (Code Civ. Proc., § 1021.5) for serving as a private attorney general and securing an injunction that principally benefits the public; and

F. Any other relief deemed just and proper.

DATED: November 12, 2026

Respectfully submitted,

/s/ Jason Harrow

Matthew J. Platkin*
Angela Cai*
Aaron E. Haier*
Conor Bradley*
PLATKIN LLP
One Gateway Center
Suite 1150
Newark, NJ 07102
mplatkin@platkinllp.com
(973) 561-1951

Tyler Whitmer
(Cal. Bar No. 248192)
Vivian Dong*
William Conlon*†
LEGAL ADVOCATES FOR SAFE SCIENCE AND TECHNOLOGY
1629 K St., NW Suite 300
Washington, DC 20006
tyler@lasst.org
(202) 204-2248

Ben Rashkovich*
LEGAL ADVOCATES FOR SAFE SCIENCE AND TECHNOLOGY
125 Park Avenue, Floor 25
New York, NY 10017
ben@lasst.org
(914) 330-4672

Jason Harrow
(Cal. Bar No. 308560)
GERSTEIN HARROW LLP
12100 Wilshire Blvd., Suite 800
Los Angeles, CA 90025
jason@gerstein-harrow.com
(323) 744-5293

Charles Gerstein*
Annelise Ernst*^
GERSTEIN HARROW LLP
1319 F Street NW, Suite 301
Washington, DC 20004
charlie@gerstein-harrow.com
(202) 670-4809

Counsel for Plaintiffs

** pro hac vice applications forthcoming*

† Admitted only in Massachusetts. Not a member of the D.C. Bar.
Practice supervised by Tyler Whitmer pursuant to D.C. Court of Appeals Rule 49(c)(8).

^ Admitted only in Louisiana. Not a member of the D.C. Bar.
Practice supervised by Charles Gerstein pursuant to D.C. Court of Appeals Rule 49(c)(8).