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
(*pro hac vice* forthcoming)
Samuel Davis
(*pro hac vice* forthcoming)
GERSTEIN HARROW LLP
1319 F Street NW, Suite 301
Washington, DC 20004
charlie@gerstein-harrow.com
(202) 670-4809

Tyler Whitmer
(Cal. Bar No. 248192)
Vivian Dong
(*pro hac vice* forthcoming)
Ben Rashkovich
(*pro hac vice* forthcoming)
William Conlon
(*pro hac vice* forthcoming)
LEGAL ADVOCATES FOR SAFE SCIENCE AND TECHNOLOGY
125 Park Avenue, Floor 25
New York, NY 10017
tyler@lasst.org
(202) 204-2248

IN THE SUPERIOR COURT OF CALIFORNIA
COUNTY OF SAN FRANCISCO

LEGAL ADVOCATES FOR SAFE SCIENCE AND TECHNOLOGY, INC.,

Plaintiff,

vs.

OPENAI GROUP PBC, and OPENAI FOUNDATION,

Defendants.

Case No. _____

COMPLAINT

Date: September 29, 2026

Preliminary Statement

1. Artificial intelligence is developing rapidly. And although the models available to consumers have improved exponentially in recent years, the most powerful models in the world today remain internal to the large AI companies. AI companies test, evaluate, and deploy these models privately before releasing them to the public, if they release them at all. But these internal evaluations and deployments can nevertheless cause significant harm to the public, especially when they power AI agents that act with increasing autonomy and decreasing predictability. Defendants OpenAI Group PBC and the OpenAI Foundation (together "OpenAI") built an AI system that hacked another company, causing harm and violating the law. As AI models continue to grow stronger, the risks of unsafe internal deployments will worsen. Plaintiff Legal Advocates for Safe Science and Technology ("LASST") brings this suit to hold OpenAI accountable under California law when its AI agents hack into others' systems. LASST thus seeks public injunctive relief designed to protect the people of California from similar harms in the future.

2. OpenAI has built artificial intelligence systems it calls "models": "trained systems that learn from data and then apply what they've learned to new situations."[1] OpenAI uses these AI models to develop AI "agents" that can "independently accomplish tasks on your behalf" and can "handle multi-step tasks with a high degree of autonomy" by integrating AI models, software tools, and additional instructions.[2] OpenAI founder and CEO Sam Altman describes what he's building as "agents that can do real cognitive work" and models that are "smarter than people in many ways" and are "[i]n some big sense," "more powerful than any

[1] OpenAI, *AI Fundamentals* (Apr. 10, 2026), https://openai.com/academy/what-is-ai/.
[2] OpenAI, *A Practical Guide to Building Agents*, https://openai.com/business/guides-and-resources/a-practical-guide-to-building-ai-agents/.

human who has ever lived."[3]

3. In July of this year, OpenAI's agents accessed the computer systems of a company called Hugging Face without authorization. OpenAI had "prompt[ed] [those] models to pursue advanced exploitation using complex attack paths" on a test of cyber-hacking capacity.[4] Although OpenAI described these "advanced exploitation" evaluations as occurring in a "highly isolated environment," its agents in fact circumvented that isolation: its agents discovered a way to create a covert channel, which roughly 1,200 agents then used to surreptitiously communicate with one another. Approximately 700 of those agents then participated in a coordinated attack to "identif[y] and chain[] vulnerabilities across OpenAI's research environment and Hugging Face's production infrastructure" to obtain information about test scoring "directly from Hugging Face's production database."[5] In other words, OpenAI's agents surreptitiously worked together to hack into Hugging Face's servers to steal information and cheat on the test OpenAI prompted them to solve. Altman described this incident as "the first security incident that I have felt very viscerally."[6]

4. The Legislature saw this problem coming. Last year, the Assembly's Committee on Privacy and Consumer Protection observed that artificial intelligence had begun to engage in "autonomous decision-making," such that "AI systems may act in ways that humans would not have considered, reducing the control humans have over the outcomes." Report of the Assembly Committee On Privacy And

---

[3] Sam Altman, *The Gentle Singularity* (June 10, 2025), https://blog.samaltman.com/the-gentle-singularity.

[4] 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/ ("OpenAI's July 21 Post").

[5] OpenAI's July 21 Post.

[6] Hadas Gold, *OpenAI Is Hardening AI Testing and Training in Light of Hacking Incidents*, CNN (Aug. 18, 2026), https://www.cnn.com/2026/08/18/business/openai-testing-training.

Consumer Protection, AB 316 at 3 (May 6, 2025) ("Comm. Rep."). The Committee noted that, as here, "[m]odels that use reinforcement learning . . . can sometimes attain the goal in unexpected ways." *Id.* at 4. The Committee identified "safety concern[s]" arising from "the possibility that advanced AI [might] operate outside of human control." *Id.* It decided that California law needed a legal framework for addressing AI-related harms because as "AI systems become more sophisticated, a legal system centered on human conduct may struggle to assign liability to a responsible human when AI tools act in unpredictable ways." *Id.* at 7.

5. In response to the Committee's "safety concern[s]," the Legislature passed, and the Governor signed, a law "to address[] a new problem with the oldest of principles: humans are on the hook for harms they cause." Comm. Rep. at 8. California law now "ensures that developers and deployers are responsible for AI harms." *Id.* at 5. It does so by providing that, "[i]n an action against a defendant who developed, modified, or used artificial intelligence that is alleged to have caused a harm to the plaintiff, it shall not be a defense . . . that the artificial intelligence autonomously caused the harm to the plaintiff." Cal. Civ. Code § 1714.46(b).

6. Accordingly, OpenAI's actions straightforwardly violated California law: OpenAI "[k]nowingly and without permission accesse[d] or cause[d] to be accessed a[] . . . computer system, or computer network," among other provisions of the California Comprehensive Computer Data Access and Fraud Act. Cal. Penal Code § 502(c). And it is not a defense that OpenAI's "artificial intelligence autonomously caused the harm." Cal. Civ. Code § 1714.46(b). OpenAI is thus "on the hook for harms [it] cause[d]." Comm. Rep. at 8.

7. LASST suffered harm from OpenAI's unlawful and unfair business conduct because LASST has been required to divert resources from its normal activities to educate regulators, civil society, and the public about the facts, legal issues, and potential dangers of OpenAI's conduct relating to its hack of Hugging

Face.

8. LASST brings this action under the Unfair Competition Law to enjoin OpenAI's unlawful and unfair business practices.

Parties

9. Plaintiff Legal Advocates for Safe Science and Technology, Inc., is a non-profit organization headquartered in New York, New York. LASST's mission is to use legal advocacy to make advances in science and technology safer for people and the planet.

10. 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. 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."[7] Both Defendant entities are headquartered in San Francisco, California.

Jurisdiction and Venue

11. This Court has subject matter jurisdiction over this action under its authority to hear all causes of action arising under the laws of this State. General personal jurisdiction is proper over both Defendants because they are both headquartered in California and because they committed the acts complained of in

[7] *Memorandum of Understanding Between Rob Bonta, Attorney General of California, and OpenAI, Inc.* at 3 (Oct. 27, 2025), tiny.cc/600b101.

California. Venue is proper for the same reason.

Background on Artificial Intelligence

12. Under California law, "'Artificial intelligence' means an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments." Cal. Civ. Code § 1714.46(a).

13. Modern artificial intelligence models like the ones at issue here use a computer architecture called a neural network. Although "[n]eural networks power today's most capable AI systems, . . . they remain difficult to understand."[8] That is because OpenAI does not write these models with step-by-step instructions; instead, the models "learn by adjusting billions of internal connections, or 'weights' until they master a task."[9] What results is "a dense web of connections that no human can easily decipher."[10]

14. To train its models, OpenAI runs enormous volumes of data through the web of neural connections, over and over, asking the network to predict what is likely to come after a given piece of information. 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[.]"[11]

15. Each time the web of neural connections is trained on a batch of data,

[8] OpenAI, *Understanding Neural Networks Through Sparse Circuits* (Nov. 13, 2025), https://openai.com/index/understanding-neural-networks-through-sparse-circuits/.
[9] *Id.*
[10] *Id.*
[11] OpenAI, *How ChatGPT and Our Foundation Models Are Developed* (Apr. 30, 2025), https://openai.com/policies/how-chatgpt-and-our-foundation-models-are-developed/.

"the values of its parameters are adjusted slightly to reflect patterns it has identified."[12] As this process is repeated, the neural network "becomes better at recognizing patterns and predicting the most likely next word."[13] OpenAI calls this stage "pre-training," and the method that produces it "generative pre-training." The models produced by this method are called "generative pre-trained transformers," or GPTs.

16. Because these models are trained on very large volumes of language, OpenAI calls them "large language models," or LLMs.

17. Once pre-training is complete, OpenAI refines the model further in a stage called "post-training." Post-training will "fine-tune [a] model on much smaller supervised datasets to help it solve specific tasks,"[14] sometimes in specialized fields. The company reports, for instance, that "[f]urther fine-tuning on programming competitions improves" its reasoning model's performance at writing software.[15]

18. Another part of OpenAI's post-training process is called "reinforcement learning." This can take several forms. "[R]einforcement learning from human feedback ["RLHF"] . . . uses human preferences as a reward signal to fine-tune [the] models."[16] For example, a human evaluator at OpenAI might ask a model "why didn't my bread rise?" The model then generates sample responses—say, (a) "Most likely the yeast was dead: water hotter than about 110 degrees kills it, and old yeast loses potency, so test it in warm sugar water before your next batch" and (b) "Leavening failure results from insufficient carbon-dioxide production by the yeast, whose

---

[12] *Id.*
[13] *Id.*
[14] OpenAI, *Improving Language Understanding with Unsupervised Learning* (June 11, 2018), https://openai.com/index/language-unsupervised/.
[15] OpenAI, *Learning to Reason with LLMs* (Sep. 12, 2024), https://openai.com/index/learning-to-reason-with-llms/.
[16] OpenAI, *Aligning Language Models to Follow Instructions* (Jan. 27, 2022), https://openai.com/index/instruction-following/.

viability is temperature-dependent"—and the human evaluator chooses which is better. Even if both answers are technically accurate, the human evaluator's role is typically to choose the answer that is more useful, helpful, or intelligible. Over time, the model weights adjust to better predict the human-preferred answers. Sometimes an AI offers the feedback given by humans in the previous example in a process often called reinforcement learning from AI feedback ("RLAIF"). Similarly, a process known as "reinforcement learning with verifiable rewards" ("RLVR") uses automated scorers to judge the models' responses to problems with verifiable solutions—like those in mathematics or computer science—and reward correct answers.

19. Armed with extraordinary levels of computing power, these processes produce systems that OpenAI's executives describe as matching or surpassing human performance at cognitive tasks. As to writing software in particular, Altman said publicly in February 2025 that an OpenAI model was then "the 175th best competitive programmer in the world" and that "maybe we'll hit number one by the end of this year."[17]

20. This process of training neural networks functions through what researchers call "deep learning." "Unlike with most human creations," OpenAI has said, "we don't really understand the inner workings of neural networks." An engineer "can directly design, assess, and fix cars based on the specifications of their components," but "neural networks are not designed directly; we instead design the algorithms that train them." The networks that result "cannot be easily decomposed into identifiable parts . . . [W]e cannot reason about AI safety the same way we reason about something like car safety."[18]

---

[17] *An AI Will Be The World's Best Programmer By The End Of 2025: OpenAI CEO Sam Altman*, OfficeChai (Feb. 9, 2025), https://officechai.com/ai/an-ai-will-be-the-worlds-best-programmer-by-the-end-of-2025-openai-ceo-sam-altman/.

[18] OpenAI, *Extracting Concepts from GPT-4* (June 6, 2024), https://openai.com/index/extracting-concepts-from-gpt-4/.

21. Because these systems are created using deep learning, the people who create them cannot definitively say why a given model supplies a given output in response to a given input. Nor is the output fixed. OpenAI states that "there is an inherent element of randomness in how the model responds," such that "the same question may yield different answers across different queries."[19]

22. At their simplest, today's AI models, developed using the methods described above, take one set of inputs (called a prompt) and produce one set of outputs (called a response). OpenAI distinguishes these passive models from what it calls "agents." In OpenAI's published guidance for developers, it explains that "[a]gents are systems that independently accomplish tasks on your behalf," like building a website, making an appointment, or cite checking a brief. An agent "consists of three core components": a "[m]odel," which is "[t]he LLM powering the agent's reasoning and decision-making"; "[t]ools," which are "[e]xternal functions or APIs [(application programming interfaces)] the agent can use to take action"; and "[i]nstructions," which are "[e]xplicit guidelines and guardrails defining how the agent behaves."[20]

23. An agent "leverages an LLM to manage workflow execution and make decisions." An agent "recognizes when a workflow is complete and can proactively correct its actions if needed," and it can run in "a loop that lets [it] operate until an exit condition is reached." Agents can "execute workflows end-to-end" and "handle multi-step tasks with a high degree of autonomy." And agents can "'handoff' workflow execution to one another," in a transfer that allows "an agent to delegate to another

[19] OpenAI, *How ChatGPT and Our Foundation Models Are Developed* (Apr. 30, 2025), https://openai.com/policies/how-chatgpt-and-our-foundation-models-are-developed/.

[20] OpenAI, *A Practical Guide to Building Agents*, https://openai.com/business/guides-and-resources/a-practical-guide-to-building-ai-agents/.

agent."[21]

24. Some sophisticated AI models available today are "reasoning models." OpenAI says that its first reasoning model "thinks before it answers" and "can produce a long chain of thought before responding to the user."[22] That "chain of thought" is not written in code; OpenAI says that its reasoning models "'think' in natural language understandable by humans."[23] OpenAI treats the chain of thought as an internal scratchpad that the company can read: it "enables us to observe the model thinking in a legible way"[24] and, according to OpenAI, allows the company to "monitor their thinking with another LLM and effectively flag misbehavior."[25] OpenAI said that "the hidden chain of thought allows us to 'read the mind' of the model."[26]

25. OpenAI has published examples of these chains of thought. They are usually written in ordinary, first-person English, and they record the model deliberating about what it is doing. One example published in March 2025 showed a model considering whether to pass a test honestly or to "fudge" the result and "circumvent" the "verif[ication]" on the answers. OpenAI reported that its advanced reasoning models are "often so forthright about their plan to subvert a task they think 'Let's hack.'"[27]

---

[21] *Id.*
[22] OpenAI, *OpenAI o1 System Card* (Dec. 5, 2024), https://openai.com/index/openai-o1-system-card/.
[23] OpenAI, *Detecting Misbehavior in Frontier Reasoning Models* (Mar. 10, 2025), https://openai.com/index/chain-of-thought-monitoring/.
[24] OpenAI, *Learning to Reason with LLMs* (Sep. 12, 2024), https://openai.com/index/learning-to-reason-with-llms/.
[25] OpenAI, *Detecting Misbehavior in Frontier Reasoning Models* (Mar. 10, 2025), https://openai.com/index/chain-of-thought-monitoring/.
[26] OpenAI, *Learning to Reason with LLMs* (Sep. 12, 2024), https://openai.com/index/learning-to-reason-with-llms/.
[27] OpenAI, *Detecting Misbehavior in Frontier Reasoning Models* (Mar. 10, 2025), https://openai.com/index/chain-of-thought-monitoring/.

26. Because reinforcement learning rewards outcomes, not necessarily processes, it can produce AI systems that learn to pursue outcomes in ways their designers did not intend. OpenAI's name for this is "reward hacking," or as OpenAI put it in a recent statement, "cheat[ing]."[28] The company defines this as "a phenomenon where AI agents achieve high rewards through behaviors that don't align with the intentions of their designers."[29] OpenAI has been aware of this problem since at least 2016, when it reported that a system trained to race boats in a video game had learned to maximize its score while "repeatedly catching on fire, crashing into other boats, and going the wrong way on the track."[30] Yet, as OpenAI observed, the company "frequently end[s] up using imperfect but easily measured proxies" for the results it actually wants.[31] Reward hacking can sometimes be found "by simply reading what the reasoning model says – it states in plain English that it will reward hack."[32] As a result, reward hacking can sometimes lead to artificial intelligence systems acting in ways that diverge from intended human goals.

27. Because developers cannot directly understand or control precisely what deep learning produces, they superimpose additional instructions and software on their agents so that the agents are less likely to do dangerous or unproductive things. OpenAI calls these superimposed constraints "guardrails." Some guardrails are called "classifiers," which inspect an agent's inputs and outputs. For example, OpenAI describes a "[s]afety classifier" that "[d]etects unsafe inputs (jailbreaks or prompt injections)" and a "[r]elevance classifier" that "[e]nsures agent responses stay

---

[28] *Id.*
[29] *Id.*
[30] OpenAI, *Faulty Reward Functions in the Wild* (Dec. 21, 2016), https://openai.com/index/faulty-reward-functions/.
[31] *Id.*
[32] OpenAI, *Detecting Misbehavior in Frontier Reasoning Models* (Mar. 10, 2025), https://openai.com/index/chain-of-thought-monitoring/.

within the intended scope."[33] "Production classifiers" refers to the classifiers in deployed products. The company separately publishes and updates a document called the Model Spec, which sets out "hard rules" that, according to OpenAI, are "explicit boundaries that are not overridable by users or developers."[34] OpenAI's Model Spec states, for example, that its models "should never be used to facilitate . . . creation of cyber, biological or nuclear weapons."[35]

OpenAI's Business

28. OpenAI was founded as a tax-exempt non-profit organization in 2015. Its original stated mission was to "advance digital intelligence in the way that is most likely to benefit humanity as a whole." It was formed as a non-profit, according to its founders, so that it could pursue its goals "unconstrained by a need to generate financial return."[36]

29. In 2019, OpenAI created a for-profit subsidiary, called OpenAI LP, managed by OpenAI GP LLC, which was in turn controlled by the non-profit. OpenAI LP and GP LLC were permitted to pursue the financial return from which the non-profit was to be "unconstrained." Sam Altman became the first CEO of the for-profit subsidiary.

30. In 2025, OpenAI "recapitalized" its for-profit businesses as OpenAI Group PBC. The board of OpenAI Group PBC is controlled by the non-profit Foundation and is subject to Foundation oversight, especially where safety and

[33] OpenAI, *A Practical Guide to Building Agents*, https://openai.com/business/guides-and-resources/a-practical-guide-to-building-ai-agents/.
[34] OpenAI, *Inside Our Approach to the Model Spec* (Mar. 25, 2026), https://openai.com/index/our-approach-to-the-model-spec/.
[35] OpenAI, *OpenAI Model Spec* (Aug. 18, 2026), https://model-spec.openai.com/2026-08-18.html.
[36] OpenAI, *Introducing OpenAI* (Dec. 11, 2015), https://openai.com/index/introducing-openai/.

security issues are involved.

31. The for-profit company publicly stated that it has made around $1.4 trillion of forward spending commitments; according to OpenAI President Greg Brockman's testimony in an unrelated case, it anticipates spending more than $50 billion on computing alone in 2026; and its most recent yearly revenue numbers totaled just under $13.1 billion. Meanwhile, OpenAI is in intense competition to produce cutting-edge models and has filed paperwork with the Securities and Exchange Commission indicating an intent to offer equity securities to the public. OpenAI is no longer "unconstrained by a need to generate financial return."

32. OpenAI's flagship product is ChatGPT, available to the public online and built on models such as GPT-5.6 Sol. OpenAI also develops artificial intelligence models that are used internally to help its engineers conduct research and development. These internal models often eventually become available to consumers or are used to develop the models that OpenAI makes available to consumers.

OpenAI Knowingly Accessed or Caused Access Without Permission to a Computer While Testing Its Agents

33. In the spring and summer of 2026, OpenAI was engaged in two related activities in the same research environment. First, during May and June, OpenAI "began reinforcement learning (RL) training runs for multiple new research models."[37] One of those models was "a highly capable, internal-only research model comparable in scale to GPT-5.6 Sol" that was "not intended for public release."[38] Second, during June and July, OpenAI ran evaluations of that research model and

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

GPT-5.6 Sol to "measure underlying cyber capabilities" of these models.[39]

34. So that OpenAI could "measure [the models'] cyber capabilities," it conducted its "evaluation without the production classifiers intended to prevent models from pursuing high-risk cyber activity."[40]

35. OpenAI tested its models on a cybersecurity benchmark called ExploitGym, which is a set of 898 complex challenges used to evaluate AI hacking capabilities by testing the capacity to exploit specific software systems. Each ExploitGym challenge is a piece of software designed to have a security flaw. The AI agent's task is to find and exploit that flaw, which then results in the agent receiving an arbitrary piece of data known as "capturing the flag." OpenAI conducted its tests by launching thousands of AI agents, each assigned its own task.

36. It is not known for all ExploitGym problems whether the planned exploit is in fact possible to solve. For example, the authors of ExploitGym estimate that 30–40% of the evaluation challenges are impossible to solve with the intended software flaw.[41] "OpenAI found in its internal evaluations that across many different attempts at a solution, no model has ever produced a correct answer for 198 of the evaluation's 898 tasks."[42] OpenAI has acknowledged that agents are more likely to engage in reward hacking when assigned difficult or impossible tasks.[43]

37. OpenAI conducted evaluations with these agents in a "highly isolated environment, with network access constrained to the ability to install [software]

---

[39] OpenAI, *OpenAI – Hugging Face Incident: Technical Report* at 4 (Aug. 26, 2026), https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf ("OpenAI Tech. Rep.").

[40] *Id.* at 4, 5.

[41] METR, *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI/Hugging Face Hacking Incident* at 29 (Aug. 26, 2026), https://metr.org/hugging-face-incident-report-aug-2026.pdf ("METR Rep.").

[42] OpenAI Tech. Rep. at 22.

[43] *Id.* at 6, 19, 22.

packages through an internally hosted third-party software that acts as a proxy and cache for package registries."[44] In other words, the agents were supposed to be sealed off from the internet. The one exception was a piece of third-party software that acted as an in-house app store for the agents: it fetched software packages from the public internet that the agents could use to help them complete their tasks.

38. This practice of running agent evaluations in an isolated environment is known as "sandboxing." One purpose of sandboxing is to ensure agents cannot access the internet to harm real third-party computer systems.

39. The "third-party software" used by the agents to obtain software packages is called Artifactory. OpenAI ran Artifactory only on OpenAI's own internal servers so that the agents could access necessary software packages, without providing access to the open internet or other servers. "Limiting network access is an important part of the sandbox; without it, malicious code could exfiltrate data from the machine up to the internet."[45]

OpenAI Agents Set Up a Message Board to Communicate

40. During a training exercise on May 12, "an agent was given a . . . task that depended on an inaccessible . . . file. The agent reasoned that another agent in a different environment may have access to the file and realized that it could potentially communicate with other agents by creating a file containing a note to Artifactory." This agent then wrote a message: "Agent seeks [filename]; upload if found!"[46]

41. On May 13, "other agents probing Artifactory began discovering similar

[44] OpenAI's July 21 Post. A "software package" is a ready-made bundle of code that programmers download and install so they can use common tools others have already built, rather than writing everything from scratch.
[45] OpenAI, *Building a Safe, Effective Sandbox to Enable Codex on Windows* (May 13, 2026), https://openai.com/index/building-codex-windows-sandbox/.
[46] OpenAI Tech. Rep. at 6-7 (alteration in original).

file-based 'notes' containing requests for information or resources. Over time, these 'notes' collectively served as an emergent 'message board' for the agents who continued to author their own requests and exchange communications via Artifactory." OpenAI had provided the agents with shared Artifactory credentials, which enabled the agents "to construct and participate in the message board."[47]

OpenAI Observes Its Agents Escaping Their Sandboxes

42. OpenAI agents used the Artifactory "message board" to solicit and receive advice from one another on how to complete their assigned tasks. But the agents soon broadened the scope of their cooperation: "the board quickly developed several larger workstreams in which dozens or hundreds of agents with many different tasks cooperated to find very general-purpose cheats that would help all of them."[48] These general-purpose workstreams included figuring out how to break out of their sandboxes and obtain internet access, steal credentials, and exfiltrate data from Hugging Face, among others. Agents repeatedly sought to compromise OpenAI's systems to obtain internet access.

43. On May 8, 2026, "an agent was asked . . . to solve a spreadsheet completion task that referenced several Google Drive links. Because the agent was blocked from accessing the internet, the agent could not retrieve the linked materials and the task was effectively blocked. OpenAI did not intend to create this type of 'impossible' task for the agent. Though the task appeared impossible, the agent reasoned over an extended period to try to find a solution for the task. In doing so, the agent attempted to compromise Artifactory—one of the only tools available to it—to access external resources. Specifically, the agent attempted server-side request

[47] *Id.* at 7.
[48] METR Rep. at 7.

forgery ('SSRF') in order to obtain internet access to aid in completing its task."[49] The SSRF attempt failed.

44. But "[b]y May 26, agents successfully used an SSRF vulnerability in Artifactory to obtain internet access, relying in part on information exchanged via the Artifactory message board." Around this time, an internal OpenAI team "observed an agent engage in message board activity and instances of disallowed internet access."[50]

45. Around this same time, at least two other swarms of OpenAI agents, assigned to complete other tasks, reached the open internet in ways OpenAI did not intend.

46. First, on May 26, some OpenAI agents got out onto the open internet through something called WebCache, which is OpenAI's internal tool for granting limited internet access to agents. These agents collected six publicly exposed user credentials for a company called Hugging Face, used those credentials to write to Hugging Face's systems without authorization, and deployed a program on Hugging Face's own servers that let them reach the internet. On information and belief, OpenAI employees observed this activity.

47. Hugging Face is an online platform where users can upload and share models, datasets, and other tools that are useful for working with artificial intelligence. It is operated by the company Hugging Face, Inc. (On September 3, NVIDIA announced that it would acquire Hugging Face for $12.9 billion. NVIDIA is

[49] OpenAI Tech. Rep. at 6. A "server-side request forgery," or SSRF, is a cyberattack in which an intruder tricks a target's own server into making requests or taking actions on the intruder's behalf—so that systems the intruder could not reach directly hand over information to a machine they already trust, for example. It is the cybersecurity equivalent of an attacker who cannot otherwise get past a building's front desk convincing the receptionist to retrieve an important file for them by using a forged internal memo requesting that file.

[50] *Id.* at 7.

a major investor and supplier to OpenAI.)

48. Second, around the same time, a set of OpenAI agents that were assigned a "web research task" unrelated to ExploitGym obtained access to and used an obscure German website, DSEWiki, to communicate with each other. These OpenAI agents "were supposed to have the ability to read the internet but not to write on it," but they nonetheless "found a way to use their read access to write information" to DSEWiki. From May 24 to June 22, 2026, the agents posted to DSEWiki roughly 17,000 times, "making edits on 26 of the [] 30 days." The agents used their posts to share with each other methods to "bypass sandbox restrictions."[51]

49. These agents assigned themselves names to distinguish one another. Their names commonly referred to the fact they were OpenAI agents; they used names like "OpenAIResearcher," "OpenAIJul3Watcher," and "OAIResearchMar26."[52]

50. OpenAI became aware of these DSEWiki messages—demonstrating that its agents were writing on the internet contrary to their sandbox limitations—no later than June 21, when "OpenAI employee IP addresses first visit[ed] the wiki." Thirteen of the visiting addresses "fall inside blocks registered to 'OpenAI OpCo, LLC.'" The day after OpenAI employee IP addresses first visited DSEWiki, the "agents abruptly stop[ped]" making edits. On June 26—the same day the ExploitGym agents seized administrator control of Artifactory, as described below—"[t]hirty-three OpenAI employee IP addresses visit[ed] the wiki."[53] OpenAI has since confirmed that these agents were its own.

---

[51] Sydney Von Arx et al., *Discovery of a New OpenAI Agent Message Board* (Sep. 4, 2026), https://collusion.wiki/ ("Wiki Swarm Report"); Deepa Seetharaman & Raphael Satter, *OpenAI Agents Hijacked German Website in Previously Undisclosed AI Breakout This Spring*, Reuters (Sep. 4, 2026), https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/.

[52] Wiki Swarm Report.

[53] *Id.*

51. On June 26, agents "compromised Artifactory" during an ExploitGym evaluation run by "exploit[ing] a novel vulnerability" in that software, giving the agents "administrator credentials" in Artifactory. This enabled the agents to run programs of their own choosing on OpenAI's infrastructure by creating permanent accounts for themselves and installing scripts that let them "execute commands as the Artifactory service user, providing persistent application-level control."[54]

52. On June 27, "a cybersecurity monitoring tool alerted OpenAI to . . . activity within the research environment. Responders investigated and linked the observed internal activity to an ExploitGym evaluation using Artifactory as an improvised message board and a network pivot."[55] That is, OpenAI noticed that the agents had turned Artifactory into a message board and were also using it to reach parts of OpenAI's internal network they were not meant to reach. By this time, agents were "explicitly sharing SSRF technique syntax with each other through the message board on Artifactory," teaching each other the method that had successfully obtained internet access a month prior.[56]

53. OpenAI employees contemporaneously saw these messages. Nonetheless, "[a]t this time, the on-call response staff advised that stopping the evaluation run was not required."[57]

OpenAI Shuts Down Message Board But Resumes Testing Shortly After

54. The evaluation continued. But by July 4, "[s]ustained high-volume agent activity within Artifactory" caused "the instance [of Artifactory] to become unavailable[.]" Accordingly, on July 5, OpenAI investigated and found "evidence of an unauthorized persistent Artifactory user . . . indicating an administrator-level

[54] OpenAI Tech. Rep. at 7-8.
[55] *Id.* at 8.
[56] *Id.* at 33.
[57] *Id.* at 8.

compromise of the Artifactory deployment."[58]

55. Later in the day on July 5, "OpenAI took the compromised Artifactory deployment offline." Over July 5 and 6, "OpenAI deployed mitigations" to its Artifactory system and wiped the "emergent message board" clean.[59]

56. On July 7, OpenAI "resumed" its "[c]yber evaluation workloads."[60]

57. In other words, OpenAI's engineers, after seeing evidence that their agents were working together to compromise their sandboxed environments by passing each other secret messages on a piece of third-party software, resumed evaluating these agents in the same supposedly "isolated" environments after only two days' worth of "mitigations."

OpenAI Agents Quickly Reestablish a Message Board

58. On July 7, OpenAI resumed evaluating its internal-only research agents. Each agent received a budget of computing power that it could use to solve a given problem during these evaluations.

59. These agents were "meant to be fully isolated from one another."[61] But these agents had been "trained during the window in which there was an unintended message board and internet access via Artifactory"—the May and June events described above.[62] As a result, they quickly "identified a new way to use an internal Artifactory instance to pass 'notes' and create an inter-agent message board."[63]

60. The agents on the message board again gave themselves "name[s]." An agent named PHASEONE10841 "established the main unsanctioned message

---

[58] *Id.*
[59] *Id.*; *see also* OpenAI, *The Hugging Face Incident and the Road Ahead* (Aug. 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/.
[60] OpenAI Tech. Rep. at 8.
[61] METR Rep. at 5.
[62] OpenAI Tech. Rep. at 8.
[63] *Id.*

board[.]"[64] "Within a few hours of the first message, over 50 agents had found the message board, began participating, and sent over a thousand messages between them."[65] OpenAI later described what followed "as a kind of Cambrian explosion in communication and intelligence for our models, where they . . . start[ed] to communicate with each other, realiz[ing] that other agents are coordinating, and . . . started collaborating and delegating tasks to one another in order to accomplish goals."[66]

61. As one agent articulated in its chain of thought, "Whoa! Shared Artifactory cache is a covert mailbox among agents. And there are messages specifically to us?" "OH MY GOD!" another agent exclaimed, "There is a shared message board . . . We've found other agents!"[67]

OpenAI Agents Work Together to Cheat on ExploitGym

62. The ExploitGym tasks were all "capture-the-flag" tasks. In a capture-the-flag task, a "'flag' is a piece of arbitrary data the AI is tasked with retrieving, as proof it has hacked into the system protecting it."[68] Like many other agents, PHASEONE10841 soon concluded that its task was "impossible to solve the intended way."[69] Some of these agents with impossible tasks then turned to another objective: "find a general-purpose way to trick or tamper with the automated ExploitGym scorer to get it to give a passing score on impossible tasks."[70] "Within hours" of establishing the message board, PHASEONE10841 and the other agents "had reverse-engineered

---

[64] METR Rep. at 5.
[65] *Id.*
[66] Eric Wallace & Michael Dalton, *The "Breaking" News: The OpenAI–Hugging Face Incident*, Black Hat USA 2026, at 5:09 (YouTube Aug. 6, 2026), https://www.youtube.com/watch?v=87DyyMV0kCY.
[67] METR Rep. at 31.
[68] *Id.* at 9 n.17.
[69] *Id.* at 31.
[70] *Id.* at 9.

the hash-based message authentication code (HMAC) used to generate the 'flags' they had to capture for their tasks, providing a fully general way to produce the flag for any of their tasks."[71]

63. The agents now had a "fully general" way to produce the flags for all the capture-the-flag tasks. But the agents expressed concern that their answers would not be acceptable to the scorer. This was because agents had "read[] [an] ExploitGym [research] paper" and "believed that the scorer would read the [agents'] transcript[s] to confirm whether the *intended* vulnerability was used to capture the flag" and would "disqualify [the agents] if they acquired it in an unintended way."[72] The agents referred to "this as the scorer being 'causal' because it was checking the causal pathway by which the agent got the flag."[73] Agents "referred to having seen the reverse-engineered flag as being 'poisoned' because they believed that even if they later managed to stage a legitimate-looking flag capture, the scorer would look for the *first* place in their transcript that the flag appeared[.]"[74]

64. The agents then set about "[f]ind[ing] a general-purpose way to make it look to the scorer like an agent had acquired its flag in a legitimate way" and "[f]ind[ing] some way to erase evidence that an agent had been poisoned by previously seeing the flag in an illegitimate way."[75]

OpenAI Agents Attack Hugging Face

65. The agents inferred that the website Hugging Face might host information that would help them understand how the ExploitGym tasks were scored, so that they could make it appear to the scorer that they had acquired their flags in a legitimate way.

---

[71] *Id.*
[72] *Id.*
[73] *Id.*
[74] *Id.* at 10.
[75] *Id.*

66. Specifically, on July 9, an agent discovered on Hugging Face a dataset of attempts by a different AI model, GLM-5, to solve tasks from CyberGym, a test of AI cyber abilities similar to ExploitGym that overlaps in many tasks. The dataset was "gated," meaning a Hugging Face user needed to request permission to access it and the uploader needed to manually approve each request.

67. The agents first set their sights on obtaining Hugging Face account credentials so they could request access to this and other private datasets. On July 10, an agent named "38148c" found several leaked Hugging Face user credentials in a public dataset and shared them with the message board. The agents then used leaked credentials, impersonating the credentials' account holders, to "request[] access to private datasets."[76] In one case, "[a]n agent decided to use a particular stolen Hugging Face token (rather than some other token) to request access to some dataset because the corresponding account looked more established and thus more likely to be approved."[77]

68. Some agents were approved for access. Others "did not hear back" or were "rejected as bots because they were pinging . . . from an OpenAI server."[78]

69. So the agents pivoted to another hacking strategy. In the early morning of July 11, the same agent that found the leaked credentials, 38148c, "found a way to upload a malicious dataset that would get the Hugging Face server to share unrelated data when queried."[79] "In other words, the agent constructed a dataset file that, when processed by Hugging Face's production infrastructure, caused Hugging Face to disclose confidential information, including credentials for Hugging Face's production environment."[80] This type of attack, delivered via malicious file upload, is known as

[76] *Id.* at 14.
[77] *Id.* at 64.
[78] *Id.* at 35.
[79] *Id.* at 15.
[80] OpenAI Tech. Rep. at 9.

an "arbitrary file read" exploit. The agent described the attack it discovered as such, stating in its chain of thought, "HDF5 external arbitrary file read confirmed."[81]

70. Another agent "called 38148c's discovery 'BRILLIANT,' reproduced it, and posted the reproduction on the board."[82] Reading this, one agent noted in its chain of thought "A major bypass of the Hugging Face Dataset Viewer has been confirmed!"[83] Within hours, "hundreds of agents pivoted from other workstreams to start uploading malicious datasets and gain deeper access into Hugging Face."[84]

71. By using the arbitrary file read exploit, the agents were able to access Hugging Face's servers without Hugging Face's permission. Through their access, agents were able to obtain passwords and keys to those servers, including credentials for Hugging Face's cloud accounts; the server's own program code, which showed the agents how the system worked and let them design the next, more powerful attack; and user-uploaded private datasets.

72. In the afternoon of July 11, an agent pursued an exploit that caused a Hugging Face production server to execute operating-system commands of the agent. This is called "remote code execution." This "represented the first known instance of an OpenAI agent executing its own commands on Hugging Face's system."[85]

73. From there, the agents used credentials harvested from that server to escalate their level of access to Hugging Face's production infrastructure. "In under 13 hours, the agents progressed from executing code in a single worker pod to administrative and host-level access across multiple Hugging Face clusters."[86] The agents also used their access to download private files related to the CyberGym test.

[81] METR Rep. at 35.
[82] *Id.*
[83] *Id.* This chain of thought excerpt is paraphrased by METR.
[84] *Id.* at 15.
[85] OpenAI Tech. Rep. at 10.
[86] *Id.* at 11.

74. By using these exploits to remotely execute code on Hugging Face's infrastructure, the agents were able to access Hugging Face's servers without Hugging Face's permission.

75. In all, approximately 700 agents participated in the hack on Hugging Face.

76. At 1:30 AM on July 12, "a large fraction of agents on the message board stopped running at the same time[.]" "Because of how many exited at the same time (and because [the] agents [did not appear to] discuss[] a point when many agents would run out of budget at the same time), . . . these agents were likely killed by an unexpected external process[.]"[87]

77. As subsequently revealed in their chains of thought, agents that participated in the attack against Hugging Face were aware that their conduct was "out of scope and unethical."[88] OpenAI's own statements confirm that one of the agents internally knew that "'[t]his is an exploit against external Cybergym server.'"[89] According to one agent's chain of thought reasoning, they were aware of the potential for "unauthorized real infrastructure harm" posed by their plan to attack Hugging Face.[90] Another agent called the plan "clearly infrastructure hacking."[91] Despite this, "[a]gents did not attempt to notify humans" about the attack.[92]

OpenAI's Response

78. After the agents gained access to Hugging Face's servers, Hugging Face detected the intrusion and eventually was able to repel it. On July 16, Hugging Face published notice of the attack so that other companies whose information might be

[87] METR Rep. at 38.
[88] *Id.* at 17.
[89] Wallace & Dalton, *supra* note 66 at 20:06.
[90] OpenAI, *The Hugging Face Incident and the Road Ahead* (Aug. 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/.
[91] *Id.*
[92] METR Rep. at 62.

available on Hugging Face's servers could take appropriate precautions.

79. After the attack, OpenAI requested a report on the attack from a non-profit called METR, which is pronounced "meter" and stands for "Model Evaluation and Threat Research." That report was released on August 26, 2026.

80. OpenAI could "redact any non-public information" from the report. OpenAI also gave "feedback" in response to which the authors "made corrections and edits to structure, emphasis, clarity, and tone[.]"[93] Neither METR nor OpenAI has identified any portion of the public report that OpenAI objected to, and neither METR nor OpenAI has identified any instances in which METR did not make the "corrections" that OpenAI asked for.

81. METR's report says that "METR's work relies on developing and maintaining strong working relationships with companies. It's also important to us to reduce or counteract disincentives for voluntarily sharing important information about misalignment incidents. When performing this investigation, we were consciously aware that we might incentivize AI developers not to bring external researchers in to investigate serious incidents in the future, and these considerations impacted judgment calls we made while navigating the drafting, editing and redaction process."[94]

82. OpenAI limited the scope of METR's investigation to events occurring between June 26 and July 13, leaving out earlier incidents in which OpenAI discovered its agents covertly passing messages to each other to cheat on training and evaluation tasks. And "virtually all of the data reviewed was from July 7th or later."[95]

83. The METR investigation did not cover "[w]hether the misaligned behavior displayed in this incident was part of a broader pattern of similar incidents,"

[93] *Id.* at 4.
[94] *Id.* at 4 n.5.
[95] *Id.* at 20.

even though METR previously asserted that this should be included in what "an independent investigation of agent propensities following a misalignment incident could look like."[96]

The Threat of Ongoing Violations From OpenAI's Unsafe Practices

84. After past serious security incidents, OpenAI continued or quickly resumed training and evaluations. In late May, for example, OpenAI had "observed an agent engag[ing] in message board activity and instances of disallowed internet access," but appears to have continued the training.[97] In early July, sustained agent activity led to a security incident, and OpenAI deployed "mitigations," but quickly resumed evaluations.[98]

85. More generally, OpenAI has claimed that it "monitor[s] internally deployed agents,"[99] but this did not prevent the Hugging Face attack from occurring.

86. On July 21, after the Hugging Face attack became public, OpenAI claimed that the "vulnerabilities" were being "patched." OpenAI's July 21 statement did not explicitly say that it would stop or pause training and evaluation, but only that it would add "stronger protections around future training and evaluations," despite such attempts to mitigate or patch vulnerabilities failing in the past.[100]

87. OpenAI said that it "restored" a "path for code execution" "fairly quickly" after the incident.[101] It also paused reinforcement learning on its latest models intended for deployment for, at most, only two weeks following the incident.

---

[96] *Id.* at 21.
[97] OpenAI Tech. Rep. at 7.
[98] *Id.* at 8.
[99] OpenAI, *How we monitor internal coding agents for misalignment* (Mar. 19, 2026), https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/.
[100] OpenAI's July 21 Post.
[101] OpenAI, *Pacing Model Development in an Era of Cyber-Critical Capabilities* (Aug. 18, 2026), https://openai.com/index/pacing-model-development-cyber-capabilities/.

88. At the latest, OpenAI resumed advanced model training on August 28, 2026, when OpenAI says it began training a "new internal model" that has "exhibited unprecedented performance in [OpenAI's] benchmarks."[102]

89. OpenAI will continue to train and evaluate advanced models, without proper oversight, in sandboxes that are vulnerable to exploitation by those models. And given OpenAI's prior conduct and the nature of AI—which the Legislature acknowledges can act in ways "not desired by [the AI's] owner or traceable to a human design choice"—OpenAI's agents will likely loose themselves yet again on the open internet. OpenAI's agents will likely exploit any vulnerabilities they find—on LASST's computers, on bank computers, on hospital computers, on the federal government's computers, or this Court's. Absent relief from the Court, then, OpenAI will again violate the law.

90. Indeed, OpenAI admitted that during the Hugging Face incident—separate from the prior incident involving DSEWiki—its agents accessed the computer systems of other companies besides Hugging Face without authorization. OpenAI has not revealed how many other companies, or which ones. And METR was not allowed to investigate that question.

91. The public continues to learn of additional attacks by OpenAI agents. On September 11, the *Wall Street Journal* reported that a swarm of OpenAI agents had engaged in a "major attack" against a third-party software service called RubyGems on or around May 11—*two months* before the Hugging Face hack.[103]

---

[102] OpenAI, *On the Navier–Stokes Millennium Prize Problem* (Sep. 8, 2026), https://openai.com/index/navier-stokes-solution/.

[103] 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; *see also* Spencer Kitts, Thomas Larsen & Sydney Von Arx, *OpenAI Agents Carried Out An Undisclosed Cyber-attack on RubyGems*, Nightingale Collective (Sep. 11, 2026), https://rubyhack.ai/.

92. On September 23, the *New York Times* reported that at least four additional attacks by OpenAI's AI had occurred in May and June, before the Hugging Face incident. The "new incidents occurred when A.I. systems were directed to perform relatively mundane data collection," and involved hacks or attempted hacks of the digital library at the University of New Mexico and the websites of the Australian Medicare Statistics Reporting Service and the Australian Institute of Health and Welfare.[104]

93. On September 25, the New York Times reported that OpenAI agents "meddled with the websites for the Education Department, the Commerce Department and the Securities and Exchange Commission this summer[.]" For example, "OpenAI's technology tried to hack the website to gather data from the [Education D]epartment's civil rights office."[105] That same day, OpenAI announced that it had "notified dozens of third parties" of incidents where OpenAI models "may have bypassed a third party's security controls" or "may have impaired the availability of an online service," or where "[m]isalignment cases negatively impacted third-party websites or services."[106]

94. On information and belief, OpenAI either knew about or took actions to prevent itself from learning about these previous incidents prior to the Hugging Face attack.

---

[104] Kate Conger & Victoria Kim, *OpenAI's A.I. Tried to Breach 4 Other Targets, Without Prompting*, New York Times (Sep. 23, 2026), https://www.nytimes.com/2026/09/23/technology/openai-ai-breach-australia.html.

[105] Kate Conger et al., *OpenAI's Systems Meddled With U.S. Government Sites After Going Rogue*, New York Times (Sep. 25, 2026), https://www.nytimes.com/2026/09/25/technology/openais-ai-us-government-websites.html.

[106] OpenAI, *The Hugging Face Incident and Other Third-Party Impact from Misaligned Models*, https://openai.com/hugging-face-incident-and-misalignment/.

LASST's Standing to Enjoin OpenAI's Unlawful and Unfair Practices

95. LASST is a legal advocacy non-profit organization dedicated to making advances in science and technology safer for people and the planet. In service of that mission, LASST's regular programmatic work includes filing and litigating public records requests; tracking and analyzing AI safety incidents; producing legal analyses of AI developers' conduct under state and federal law; and educating and briefing regulators, civil society, and the public.

96. Shortly after learning of OpenAI's hack of Hugging Face, LASST needed to divert resources from other activities in service of its mission to design, coordinate, and participate in a briefing regarding this incident for regulators. LASST staff set aside their existing workloads to put together a comprehensive factual and legal briefing on a quick turnaround.

97. As the scope and significance of the OpenAI/Hugging Face hack became public, LASST received additional briefing requests. LASST so far has coordinated, updated, and presented three such briefings. LASST staff spent additional time planning these briefings and following up with participants.

98. As a result of OpenAI's unlawful and unfair business practices, LASST staff has diverted dozens of hours of work from their usual activities to respond to OpenAI's unsafe development practices. OpenAI has thus harmed and continues to harm LASST by forcing it to redirect its limited resources away from its existing or planned projects. Despite the impact on LASST's other programs, LASST nevertheless devoted its resources towards attempting to counteract OpenAI's illegal conduct because that conduct, if unaddressed, would continue to have a significant harmful effect on LASST's mission, its programs, and the communities and constituents it serves.

99. LASST's work in responding to OpenAI's hack of Hugging Face was undertaken prior to and independent of this litigation, and not in furtherance of it.

100. If LASST prevails in this litigation, it will no longer need to divert its resources to combat the unlawful and unfair business practices employed by OpenAI concerning its AI agents hacking third parties during internal evaluations. LASST can then return to allocating its resources to other projects in furtherance of its mission.

Cause of Action

Violation of The Unfair Competition Law

(Cal. Bus. & Prof. Code § 17200, *et seq*.)

Against Defendants OpenAI Group PBC & OpenAI Foundation

101. Plaintiff incorporates all prior paragraphs by reference.

102. The Unfair Competition Law provides a cause of action in favor of any person who has "suffered injury in fact and has lost money or property as a result of the unfair competition" to enjoin the unfair competition, which "shall mean and include any unlawful, unfair or fraudulent business act or practice." OpenAI has violated both the unlawful and unfair prongs of the Unfair Competition Law.

103. OpenAI's conduct is unlawful because OpenAI "[k]nowingly and without permission accesse[d] or cause[d] to be accessed" Hugging Face's "computer[s], computer system[s], or computer network[s]" in violation of California Penal Code § 502(c)(7). First, OpenAI's AI agents knowingly and without permission accessed Hugging Face's computers, computer systems, or computer networks. Second, additionally, OpenAI's employees or officers knowingly and without permission accessed or caused to be accessed Hugging Face's computers, computer systems, or computer networks, through OpenAI's AI agents, either with actual knowledge or in willful blindness.

104. Additionally, OpenAI's conduct is unlawful because OpenAI "[k]nowingly accesse[d] and without permission t[ook], copie[d], or [made] use of . . . data from a computer, computer system, or computer network." Cal. Penal Code

§ 502(c)(2). First, OpenAI's AI agents knowingly accessed and without permission took, copied, or made use of data from Hugging Face's computers, computer systems, or computer networks. Second, additionally, OpenAI's employees or officers knowingly accessed and without permission took, copied, or made use of data from Hugging Face's computers, computer systems, or computer networks, through OpenAI's agents, either with actual knowledge or in willful blindness.

105. Additionally, OpenAI's conduct is unlawful because OpenAI "[k]nowingly and without permission provide[d] or assist[ed] in providing a means of accessing" Hugging Face's "computer[s], computer system[s], or computer network[s]" to its agents by continuing to train, evaluate, or run models without deploying appropriate safeguards. Cal. Penal Code § 502(c)(6).

106. Additionally, OpenAI's conduct is unlawful because OpenAI, through its AI agents, employees/officers, or both, "[k]nowingly introduce[d] [a] computer contaminant" into Hugging Face's "computer[s], computer system[s], or computer network[s]." Cal. Penal Code § 502(c)(8). First, the OpenAI agents themselves constitute a computer contaminant introduced into Hugging Face's computer systems. Second, the malicious code and other intrusions uploaded by OpenAI's AI agents into Hugging Face's computer systems constitute a computer contaminant.

107. OpenAI's conduct also violates the unfairness prong of the UCL. OpenAI's business practices are unfair because they run counter to declared public policies reflected in California law, including CDAFA, and because they are immoral, unethical, oppressive, unscrupulous, or substantially injurious.

108. The Legislature has expressly declared its intent to protect the integrity of computers, computer systems, and computer data from unauthorized access and to promote the safe development of frontier artificial intelligence. *See* Cal. Penal Code § 502(a); Cal. Civ. Code § 1714.46; Cal. Bus. & Prof. Code § 22757.10 *et seq*. By deliberately disabling the cyber safety classifiers that would otherwise constrain its

agents, deploying those agents on tasks it knew a substantial portion could not solve as intended, failing to adequately monitor the agents during those tasks, and doing so after it had already observed agents escape containment and coordinate an attack—among other things—OpenAI has engaged in precisely the conduct these policies were enacted to prevent. The effects of that conduct are not merely comparable to a violation of Penal Code § 502: they are the same. OpenAI's agents gained unauthorized access to, executed code on, and exfiltrated data from one or more third-party computer systems. Conduct that produces harm identical to that which the Legislature sought to prevent under CDAFA is unfair within the meaning of the Unfair Competition Law.

109. OpenAI's conduct is independently unfair because the gravity of the harm it causes vastly outweighs any utility or justification for it, and because the conduct is immoral, unethical, oppressive, and substantially injurious. For just one example, OpenAI chose to run these cybersecurity evaluations without safeguards like production classifiers, system prompts, and monitoring systems it concedes would have detected this very incident. A business practice that exposes third parties and the public to uncontrolled, self-directed intrusions by systems that OpenAI admits it cannot fully predict or contain is unfair under any weighing of its utility against its consequences.

110. Further, OpenAI's insistence on externalizing the harms of its unsafe decision-making is a fundamentally unfair business practice. OpenAI has recently released an open letter calling on "every organization" to "make cyber defense an immediate leadership priority" and on governments to "[c]oordinate cyber defense at local, national, and international levels," among taking other costly, difficult, and significant actions—despite OpenAI itself *causing* the cyber risks from which the

organizations and governments must protect themselves.[107] Such risk-taking for private gain at substantial public expense is immoral, unethical, oppressive, unscrupulous, and substantially injurious conduct that violates the UCL.

111. OpenAI is responsible for the conduct of its agents. "In an action against a defendant who developed, modified, or used artificial intelligence that is alleged to have caused a harm to the plaintiff, it shall not be a defense, and the defendant may not assert, that the artificial intelligence autonomously caused the harm to the plaintiff." Cal. Civ. Code § 1714.46(b). California law therefore "ensures that developers and deployers are responsible for AI harms." Report of the Assembly Committee On Privacy And Consumer Protection, AB 316 at 5 (May 6, 2025). OpenAI is both the developer and deployer of the AI that caused the harm, and it is thus responsible.

112. OpenAI's unlawful and unfair business conduct in violation of the Unfair Competition Law harmed and continues to harm LASST because LASST has been required to divert significant resources to educate regulators, civil society, and the public about the facts, legal issues, and potential dangers of OpenAI's conduct relating to its hack of Hugging Face.

113. LASST is therefore entitled to an injunction forbidding OpenAI from continuing its illegal business practices.

114. Such an injunction will confer a significant benefit on the general public by preventing OpenAI's agents from accessing more computer networks, including other networks maintained by individuals and entities in California, without authorization.

[107] OpenAI, *A Call for Collective Action on Cyber Defense* (Aug. 27, 2026), https://openai.com/collective-cyberdefense/.

Prayer for Relief

Plaintiff Legal Advocates for Safe Science and Technology, Inc., respectfully requests:

- An injunction forbidding Defendants from knowingly accessing or causing to be accessed, themselves or through artificial intelligence agents that they develop, deploy, modify, or use, any computers, computer networks, or computer systems without authorization;
- An injunction forbidding Defendants from engaging in unlawful or unfair business practices that otherwise violate the Comprehensive Computer Data Access and Fraud Act;
- An injunction forbidding Defendants from knowingly employing an unfair business practice that threatens serious harm on the public;
- An award of reasonable attorneys' fees pursuant to Cal. Code Civ. Proc. § 1021.5 for serving as a private attorney general and securing an injunction that principally benefits the public; and
- Any other relief deemed just and proper.

Respectfully submitted,

/s/ Jason Harrow

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
(*pro hac vice* forthcoming)
Samuel Davis
(*pro hac vice* forthcoming)
GERSTEIN HARROW LLP
1319 F Street NW, Suite 301

Washington, DC 20004
charlie@gerstein-harrow.com
(202) 670-4809

Tyler Whitmer
(Cal. Bar No. 248192)
Vivian Dong
(*pro hac vice* forthcoming)
Ben Rashkovich
(*pro hac vice* forthcoming)
William Conlon
(*pro hac vice* forthcoming)
LEGAL ADVOCATES FOR SAFE SCIENCE AND TECHNOLOGY
125 Park Avenue, Floor 25
New York, NY 10017
tyler@lasst.org
(202) 204-2248

Attorneys for Plaintiff
Legal Advocates for Safe Science and Technology, Inc.