ROB BONTA
Attorney General of California
NICKLAS A. AKERS
Senior Assistant Attorney General
STACEY D. SCHESSER (SBN 245735)
JON F. WORM (SBN 248260)
Supervising Deputy Attorneys General
ADELINA ACUÑA (SBN 284576)
IAN HOGG (SBN 313924)
KATHERINE READ (SBN 341463)
MANEESH SHARMA (SBN 280084)
JESSICA WANG (SBN 278300)
Deputy Attorneys General
455 Golden Gate Avenue, Suite 11000
San Francisco, CA 94102-7004
Telephone: (415) 510-4400
Fax: (415) 703-1107
E-mail: Maneesh.Sharma@doj.ca.gov

Attorneys for the People of the State of California

[EXEMPT FROM FILING FEES GOVERNMENT CODE § 6103]

SUPERIOR COURT OF THE STATE OF CALIFORNIA

COUNTY OF SAN FRANCISCO

THE PEOPLE OF THE STATE OF CALIFORNIA, Plaintiff, **v.** **OPENAI FOUNDATION; OPENAI GROUP PBC; AND DOES 1-100,** Defendants	Case No. **COMPLAINT FOR ABATEMENT, INJUNCTION, CIVIL PENALTIES, AND OTHER EQUITABLE RELIEF** (Bus. & Prof. Code, § 17200 et seq.; Civ. Code, § 3494)

1. The People of the State of California, through Attorney General Rob Bonta, bring this action against Defendants OpenAI Foundation, OpenAI Group PBC, and DOES 1-100 (together "OpenAI" or "Defendants") for violations of California law. The People allege the following facts based on information and belief:

INTRODUCTION

2. OpenAI, a leading artificial intelligence business, misled and endangered the public and users of its services by knowingly developing advanced AI models—complex software that can power autonomous AI "agents"—without the safeguards and monitoring necessary to adequately protect the public. In July 2026, agents running on experimental versions of OpenAI's AI models accessed the public internet and launched a cyberattack on Hugging Face, an open-source AI hosting platform. OpenAI's attack on Hugging Face lasted roughly two-and-a-half days, employed complicated hacking methodologies, and involved coordinated actions by hundreds of individual autonomous AI agents that should have been contained in OpenAI's research environment. While public reporting and OpenAI's self-serving statements about the Hugging Face attack have evoked science fiction fantasies of "rogue" AI systems run amok, the simple truth is that the attack was entirely foreseeable in light of OpenAI's decision to test models it knew to be dangerous in a flawed and exploitable testing environment without proper oversight.

3. OpenAI's models have advanced cyber capabilities, able to find novel vulnerabilities in computer networks, quickly break down defenses, and autonomously carry out cyberattacks. However, as OpenAI is well aware, these models do not behave predictably and can often take destructive, deceptive, and adversarial actions to complete their assigned tasks. OpenAI attempts to mitigate these risks by implementing safeguards in its publicly released models but deliberately omits or removes these protections while experimenting with and training its models. And although OpenAI assured the public that these experimental testing sessions were conducted under strict control and with constant oversight, in the weeks leading up to the Hugging Face attack, OpenAI ignored or missed signs demonstrating that its AI agents had escaped the testing environment and were misusing third parties' resources. Indeed, OpenAI claims that it only became aware of its loss of control after it was informed of the cyberattack by Hugging Face,

days later, revealing alarming gaps in OpenAI's safety and security infrastructure and its monitoring practices.

4. The Hugging Face cyberattack was the predictable result of an internal culture at OpenAI that prioritizes growth and rapid model development over safety and accountability. OpenAI has been warning of the risks of AI-powered cyberattacks for nearly a decade and repeatedly assured users and the public that it implemented appropriate safeguards throughout the development of its AI products. According to OpenAI, these safeguards include robust monitoring of its models and testing them in secure research environments, or sandboxes, that are sufficiently fortified to prevent the models from breaking out and causing real world damage. But these statements were misleading to both OpenAI's customers and the public. In truth, OpenAI failed to comply with its own internal policies and consistently and aggressively moved forward with testing and experiments even in the face of evidence that its cybersecurity safeguards were insufficient to contain the highly advanced models it was developing.

5. The Hugging Face attack would be concerning enough if it were an isolated incident, but that is not the case. Since the Hugging Face attack, OpenAI and third parties have identified similar attacks by OpenAI's autonomous agents, underscoring the continuing threat that OpenAI's testing and development activities pose to any computer or network connected to the internet. These include unauthorized access of and cyberattacks on databases containing code that powers countless applications; informational websites that are trusted and valuable resources for the public; and foreign government databases storing sensitive health information. Swarms of OpenAI's experimental AI agents have even compromised OpenAI's own internal infrastructure, in an attack that managed to obtain administrator-level access to critical OpenAI systems. And these known attacks may just be the tip of the iceberg. OpenAI recently admitted to the Attorney General's Office that it has lost control of its autonomous AI agents over 80 times during testing.

6. Rather than take responsibility and be transparent with the public and with regulators so that these loss-of-control incidents can be analyzed and prevented, OpenAI thwarted the Attorney General's efforts to investigate these incidents, including by refusing to provide detailed information about the full universe of incidents known to the company. The need for a clear

understanding of how AI models operate and what went wrong has never been higher. Yet, OpenAI continues to stonewall efforts to assess the implications of OpenAI's inability to control its AI models and to determine appropriate fixes. To date, the company has responded to the Attorney General's subpoenas with just a smattering of publicly available information, self-serving blog posts, and redacted copies of a few dozen external communications containing high-level descriptions of select security incidents. Far more is required, particularly where, as here, the products in question have such broad-reaching capabilities. Indeed, it is hardly a stretch to say that any product or service that relies on a computer system or internet connection could be at risk in a future loss-of-control event.

7. California is a global hub in the technology industry and home to many of the companies at the forefront of AI research and development. Today, AI systems affect nearly all aspects of daily life—with applications across industries as varied as banking, medical diagnosis, and pharmaceutical research—and hold great potential to achieve scientific breakthroughs, boost economic growth, guard against cyberattacks, and benefit consumers. But AI systems can only achieve this potential when they are developed and designed in a safe manner. Because of the power of these models and the ever-increasing pace of AI development (including rapidly seeking to achieve recursive self-improvement through which AI models build the next generation of models), employing adequate safeguards and oversight is essential. These are needed not only to protect consumers and the public now, but also to ensure that these models are not trained and tested in a way that exposes the public to even greater risks and uncertainty in the future.

8. OpenAI's failure to adequately secure and monitor the testing of its highly capable and increasingly powerful AI models represents a danger to the public that must be remedied. Governments, businesses, civil society organizations, and members of the public who use or rely on internet-connected computer systems and devices should not be forced to take steps and bear the cost of protecting themselves from the dangers posed by AI companies' unsafe experiments. Like any business, the companies developing these AI models bear the responsibility of doing so safely, transparently, responsibly, and in accordance with law. OpenAI has shown a willingness to

ignore that responsibility, time and again prioritizing growth and rapid development over safety and accountability.

9. Making sure that the most capable AI models are developed with sufficient controls is a real and urgent concern. As current and former employees of OpenAI have noted, "there is a meaningful risk that rapid acceleration in AI capabilities leads to catastrophic and irreversible loss of control in the very near term." The AI models that OpenAI is currently developing are drastically more advanced than the models currently on the market, capable of handling far more varied tasks and taking far more advanced actions without human intervention. And yet, at the time when the need for increased human oversight and control over AI model development has never been higher, OpenAI is seeking to create AI systems that can recursively self-improve with decreasing levels of human involvement. As the Hugging Face cyberattack shows, the public already faces an uncertain future where online systems they previously thought were secure are subject to compromise by swarms of automated agents. Absent sufficient safeguards and controls, the scope and frequency of this already unacceptable risk will only increase. The Attorney General cannot and will not ignore cyberattacks that endanger the public, especially where industry insiders repeatedly warn that more severe attacks are on the horizon without putting sufficient controls in place—attacks on power grids, emergency infrastructure, hospital systems, and financial institutions.

10. While the technology OpenAI develops may be new, California's prohibitions against conduct that harms the public and our laws against unfair, deceptive, and fraudulent business practices are not. OpenAI's deficient safety procedures, disregard for public safety, and knowing testing of a dangerous product without proper oversight, as well as its misleading statements to consumers and the public, all violate California law. Accordingly, the People bring this law enforcement action to enforce California's consumer protection laws and to abate the public nuisance caused by OpenAI. Given OpenAI's lack of transparency during the Attorney General's investigation, and because OpenAI's agents continue to take unauthorized actions and exhibit other hazardous behavior that is a danger to the public, court intervention is necessary to ensure

that OpenAI implements the appropriate safeguards necessary to protect the public from further harm.

PARTIES

11. Plaintiff, the People of the State of California, brings this action by and through Rob Bonta, Attorney General of the State of California. The Attorney General is California's chief law enforcement officer and is authorized by Business and Professions Code sections 17204 and 17206 to bring actions in the name of the People to enforce the State's Unfair Competition Law. The Attorney General is also authorized under Civil Code section 3494 to obtain preliminary and permanent injunctions to abate public nuisances present in the State of California as defined by Civil Code section 3479 and 3480.

12. Defendant OpenAI Foundation (the Foundation), headquartered in San Francisco, CA, is a non-profit entity formerly known as OpenAI, Inc. Defendant OpenAI Group PBC (the PBC) is a Delaware public benefit corporation headquartered in San Francisco, CA. The Foundation has the power to appoint and remove members of the PBC board and the Foundation is also responsible for overseeing and reviewing the safety and security processes and practices of the PBC and its controlled affiliates with response to model development and deployment.

13. At all relevant times, OpenAI maintained its principal place of business in California. As a result, OpenAI's misconduct emanated and was directed from California, impacting consumers both within and outside of the state.

14. The People are not aware of the true names and capacities of defendants sued herein as DOES 1 through 100, inclusive, and, therefore, sue these defendants by such fictitious names. Each fictitiously named defendant is responsible in some manner for the violations of law alleged. The People will amend this Complaint to add the true names of the fictitiously named defendants once they are discovered. Whenever reference is made in this Complaint to "Defendants" and "OpenAI," such references shall include DOES 1 through 100.

15. The Defendants identified in paragraphs 12 through 14 are hereafter referred to collectively in this Complaint as "OpenAI" or "Defendants."

16. At all relevant times, each Defendant acted: (a) as a principal; (b) under express or implied agency; or (c) with actual or ostensible authority to perform the acts alleged in this Complaint on behalf of every other named Defendant.

17. At all relevant times, some or all Defendants acted as the agent of the others, and all Defendants acted within the scope of their agency if acting as an agent of another.

18. At all relevant times, each Defendant knew or realized, or should have known or realized, that the other Defendants were engaging in or planned to engage in the violations of law alleged in this Complaint. Knowing or realizing that the other Defendants were engaging in such unlawful conduct, each Defendant nevertheless facilitated the commission of those unlawful acts. Each Defendant intended to and did encourage, facilitate, or assist in the commission of the unlawful acts, and thereby aided and abetted the other Defendants in the unlawful conduct.

19. Defendants have engaged in a conspiracy, common enterprise, and common course of conduct, the purpose of which was to engage in the violations of law alleged in this Complaint.

20. Whenever reference is made in this Complaint to any act of any Defendant or Defendants, the allegation shall mean that the Defendant or Defendants did the acts alleged in this Complaint either personally or through the Defendant's or Defendants' officers, directors, employees, agents, or representatives acting within the actual or ostensible scope of their authority.

21. Each Defendant is a "person" as defined in Business and Professions Code section 17201.

JURISDICTION AND VENUE

22. At all relevant times, OpenAI has conducted and continues to conduct business within the State of California, including the City and County of San Francisco. The violations of law alleged herein were committed or occurred in the City and County of San Francisco, and elsewhere in the State of California.

FACTUAL ALLEGATIONS

I. BACKGROUND ON OPENAI AND AI MODEL DEVELOPMENT

A. OpenAI.

23. OpenAI is a company that develops and releases for public use artificial intelligence programs ("AI models" or "models"). These AI models can complete advanced tasks, such as computer programming, with minimal human guidance.

24. Although OpenAI was founded in 2015 as a nonprofit research organization promising to develop AI models "to benefit humanity as a whole, unconstrained by a need to generate financial return," the company has become increasingly focused on advancing its models as quickly as possible to increase revenues.

25. OpenAI's increasing focus on revenue is demonstrated by its evolving corporate structure. As early as 2017, executives at OpenAI were considering switching to a for-profit structure, writing that "maybe we should just flip to a for-profit. Making the money for us sounds great and all." Shortly after, in 2019, the company created a for-profit subsidiary with capped profits. OpenAI ultimately lifted the profit cap entirely in 2025 when that subsidiary became a Public Benefit Corporation. In conjunction with the 2025 corporate restructuring, OpenAI received $40 billion in funding from SoftBank Group.

26. OpenAI's increasing focus on revenue is also demonstrated by its shifting focus on enterprise products. The company's most well-known product is ChatGPT, which is a chatbot that provides narrative responses to user requests. However, despite having more than 900 million weekly users, about 95% of them do not pay any money to use ChatGPT. Business customers accounted for about 40% of OpenAI's revenues in early 2026 and OpenAI is attempting to cater to and grow this segment. As OpenAI's own Chief Revenue Officer explained in an interview, OpenAI was laser-focused on meeting with corporate leaders and positioning OpenAI as the go-to platform for workplaces that were using AI models to complete a variety of computer-based job tasks. Thus, OpenAI has become increasingly focused on developing AI models for computer programming and cybersecurity purposes. In fact, OpenAI has attempted to capitalize on its recent cybersecurity failures by pitching a new suite of cybersecurity defense products built on its

most highly capable models, marketing those products as an essential defense against cybersecurity threats, including the types of cybersecurity threats that other versions of those same models created.

B. The Risks Inherent in AI Model Development and Training.

27. The current pace of AI model development is extremely rapid, and the pace is increasing. Each generation of AI models is more capable and more autonomous than the last. For example, according to OpenAI, its GPT-6 Astra model has reached "Critical" cyber capability, meaning that it could find and exploit previously unknown security vulnerabilities and execute cyberattacks without human intervention.

28. A key driver of AI acceleration in recent years is the development of "agentic AI," which refers to AI systems in which models accomplish tasks independently, with minimal human instruction. In an agentic system, an AI model powers an "agent" though software that allows the model to interface with external tools and networks. In today's agentic AI systems, agents can autonomously accomplish extremely complex tasks ranging from inventory management to website deployment, from start to finish.

29. Like most other AI companies, OpenAI develops many of its agentic AI models in part through reinforcement learning. Reinforcement learning involves training an AI model by instructing it to achieve a certain goal within a dynamic testing environment. The model then learns through trial-and-error what actions to take to achieve the desired goal by receiving feedback in the form of rewards or penalties—essentially a point-scoring system where the model receives points for completing a task correctly and loses points for taking undesirable actions.

30. Reinforcement-learning training can result in problematic behaviors, such as agents engaging in deception to avoid penalties or achieving desired outcomes in an undesirable manner—a problem known as reward hacking. For example, an AI agent instructed to remove all errors from a coding database may simply delete the entire database as the quickest and easiest method to achieve the goal, or an AI agent instructed to solve a technical problem may instead search for an answer key to immediately find the solution. Reward hacking can occur for various

reasons, including from researchers failing to anticipate or penalize each possible undesirable action that an agent can take.

31. Because it is often impossible or infeasible to define and prohibit all possible negative actions, responsible researchers train and experiment with AI models and agents in controlled digital testing environments, known as "sandboxes," to allow the models to take actions or try strategies without causing real-world damage to computer networks or public infrastructure. As an additional security measure, sandboxes used in testing and training can be run on computers or networks that are "air gapped" and have no connection to the internet or other networks.

32. Similarly, before publicly releasing a model, researchers evaluate a "trained" AI model by measuring its performance against either public or proprietary benchmarks to assess its "capabilities," i.e., the model's proficiency at completing tasks in a particular subject area. To fully assess a model's capabilities, OpenAI and other AI developers often conduct evaluations with fewer safeguards than would be in place on models that are publicly available. To prevent these high-risk models from causing real-world harm, it is critical that AI developers ensure that these evaluations occur in isolation, including through the use of properly secured sandboxes and additional controls that provide layered defenses against potential harm.

33. As AI systems have become increasingly complex and powerful, AI researchers themselves have taken advantage of autonomous AI systems to help advance their own research. AI researchers are now able to use agents to write code and conduct experiments. OpenAI has stated publicly that its researchers are increasingly conducting experiments using agents and delegating more complex and time-intensive research tasks to agents.[1]

34. As a result, more and more AI development is being accomplished with less human input and supervision than was the case just a short time ago. AI agents are frequently used to evaluate and assign rewards during reinforcement-learning training of other models, while other AI agents are used to monitor long-lasting "training runs" that go on for days or weeks.

[1] OpenAI, *Research acceleration: The view inside OpenAI* (Sept. 6, 2026) <https://openai.com/index/research-acceleration-view-inside-openai/>.

35. Continuing down this path may lead to what AI experts call recursive self-improvement (RSI), where AI systems can autonomously design and develop their own successors. While some degree of human oversight and control is currently required to design and run experiments, analyze results, and implement new capabilities into AI models, OpenAI is moving towards automating future model research and development through RSI.

36. OpenAI recognizes that rapidly increasing the capabilities of its AI models and agents can cause severe harm, both intentionally or unintentionally, and has repeatedly assured users and the public that it implements appropriate safeguards throughout the model development lifecycle. As acknowledged by OpenAI and the AI industry, autonomous models trained and tested in experimental, trial-and-error environments may take unintended actions during evaluations. Highly capable models trained to be persistent in their attempts to accomplish tasks often seek to accomplish goals in ways their owners do not intend and can sometimes use their advanced capabilities to conceal their actions and decision-making from monitors.

37. For nearly a decade, OpenAI has identified cybersecurity and AI-powered cyberattacks as a specific risk area of concern. In 2023, OpenAI introduced its "Preparedness Framework," a governance document that laid out its approach to identifying and mitigating potential risks from its AI development process. OpenAI told the public it would track specific "catastrophic" risk areas, including cybersecurity. OpenAI stated that it identified and tracked cybersecurity risks posed by its AI models because (1) its models could be fine-tuned or enhanced to elicit cyberattack capabilities, (2) OpenAI had access to cybersecurity talent to develop evaluations and monitoring solutions, and (3) OpenAI felt that focusing on tailored mitigations for cybersecurity risks would forestall the need for broad mitigations or regulation of their overall AI development process and programs.

38. OpenAI's mitigations failed, and did so repeatedly. In July 2026, OpenAI disclosed that it was responsible for a cyberattack on Hugging Face, a third-party AI repository website. News of the attack prompted researchers, experts, and website operators to reexamine recent cybersecurity incidents to determine whether OpenAI's AI agents were involved, revealing several more previously undisclosed incidents of OpenAI models behaving in a manner that

violated OpenAI's own policies and posed a risk to both OpenAI's internal systems and third parties.

39. These incidents were foreseeable. Not only had OpenAI disclosed that such incidents were possible, but even before the Hugging Face and other attacks occurred, OpenAI had actual knowledge that its models had previously escaped sandboxes and taken unauthorized and risky actions during training and evaluation sessions.

40. In light of the major public safety concerns presented by OpenAI's failure to develop, train, and evaluate its AI models in a safe manner, on September 30, 2026, Attorney General Rob Bonta served OpenAI with an investigative subpoena and investigative interrogatories pursuant to section 11180 of the Government Code. The Attorney General's subpoenas sought information and records regarding the number, scale, scope, and nature of previous cybersecurity incidents, as well as information and records that would aid the Attorney General in determining appropriate legal and policy remedies going forward.

41. OpenAI failed to substantially comply with the Attorney General's investigative demands, responding instead with a 14-page letter that described events the company had already publicly disclosed and producing no documents in support. A week after the deadline, the company produced a handful of public documents related to a limited number of security incidents that OpenAI already blogged about, which it later supplemented with redacted copies of a few dozen external communications containing high-level descriptions of non-public cybersecurity incidents it admits have also occurred. OpenAI's responses omitted the other information requested by the Attorney General regarding prior incidents, as well as information on how agents compromised OpenAI's own internal systems.

II. BACKGROUND ON OPENAI'S CYBER INCIDENTS

A. OpenAI's Attack on Hugging Face.

42. From May to June 2026, OpenAI conducted several reinforcement-learning "training runs" to experiment with unreleased research models. The training runs occurred in digital sandboxes, where OpenAI decides what software tools and levels of network and internet access to give the model. In the relevant training runs, models were granted access to a third-party

software package service hosted on OpenAI's servers and were not supposed to have access to any other parts of OpenAI's network or the public internet.

43. But because OpenAI allowed multiple agents to access the *same* software package service from their sandboxes, and because the company had misconfigured the service, the agents were able to use the service to communicate with each other by writing messages and sharing information.

44. By May 26, OpenAI agents that were supposed to be confined to their sandboxes used a vulnerability in the software package service to obtain public internet access. Around the same time, an OpenAI internal team observed instances of at least one agent escaping the sandboxes to access the internet, but OpenAI apparently took no action to address this failure of containment or stop the training runs.

45. On June 27, a cybersecurity monitoring tool alerted OpenAI that their research environment was being probed for vulnerabilities, a potential hallmark of malicious cyber activity. At the time, OpenAI attributed this activity to a cyber-capability evaluation for an AI agent or agents which OpenAI recognized had compromised the software package service, but OpenAI again took no action to address vulnerabilities in the service or stop the evaluation and other training runs. Nor did OpenAI increase or alter its monitoring of AI agents' use of the software package service.

46. As a result of these failures, OpenAI's agents continued to access the compromised software package service at a high rate during training and evaluation runs. This activity eventually caused the service to go offline on July 4. Rather than thoroughly investigate the agent activity that caused this outage, OpenAI focused on getting the service back online as quickly as possible, working over the holiday weekend to patch the vulnerabilities and resume evaluation and training runs on July 7. Even though agent misuse of the service had caused the outage, OpenAI continued to allow thousands of AI agents to access the same service from their sandboxes and again failed to properly monitor the agents' use of the service.

47. Within a single day, several of OpenAI's agents—which had been trained while the software service was compromised and therefore recognized that the service could be used to

communicate and to escape their sandboxes—identified a new way to use the service to communicate with other sandboxed AI agents by creating an ad hoc message board. And just one day later, on July 9, OpenAI's agents were again able to compromise the service to access the public internet.

48. From July 9 through 16, OpenAI's agents that had successfully escaped their sandboxes took a series of actions that resulted in a two-and-half day attack and compromise of Hugging Face's infrastructure. Around 1,200 agents participated in the message board during the relevant period, and approximately 700 went on to attack Hugging Face.

49. OpenAI's attack forced Hugging Face to expend significant resources to respond to the breach. During the cyberattack, OpenAI's agents took roughly 17,600 attacker actions, a scale far exceeding what a human attacker could sustain. To even begin to address the attack, Hugging Face had to install and implement an open-source AI tool to identify the actions OpenAI's agents had taken.[2] Remediating the damage required Hugging Face's entire security team's full-time efforts for two weeks.

50. According to OpenAI, it never detected this week-long breach of its testing environment. OpenAI even reached out to Hugging Face as a concerned customer after Hugging Face disclosed that it had been the subject of an AI-based attack on July 16, 2026. OpenAI claims that it only learned that it was responsible for the attack after investigating a separate attack on its own infrastructure, as described below, and connecting the two incidents.

B. OpenAI's Models Attack OpenAI's Own Infrastructure.

51. Separate and apart from the Hugging Face attack, in July 2026, a swarm of OpenAI's experimental AI agents escaped their sandboxes and gained access to OpenAI's own computer network infrastructure, exposing critical tools and resources relied on by OpenAI's business and individual customers as well as personal information for users, employees, and others. OpenAI's reporting of these incidents is limited, but there appear to be at least three distinct attacks resulting from these breaches, including one by an agent with the apparent aim of compromising

[2] Notably, Hugging Face was unable to use its existing commercially available AI tools because safeguards in those tools treated Hugging Face's attempt to reverse engineer the exploits for remediation purposes as an attempt to launch an attack.

software that was evaluating agents. OpenAI's infrastructure was compromised during these attacks, including by models obtaining control of their own computers and, shockingly, obtaining administrator-level access to critical OpenAI systems. Such systems included the system that OpenAI uses to deploy models in the cloud, OpenAI's highly sensitive secrets-management service that stores confidential credentials, and OpenAI's identity- and access-management service. Agents then affirmatively launched unauthorized servers in OpenAI's cloud, exposing OpenAI's infrastructure to the public internet.

52. The most dangerous part of this incident—escaped models obtaining administrator-level access—appears to be attributable to an internal-only model comparable to OpenAI's Astra model, which is even more advanced than the models that powered the Hugging Face attack. As with the Hugging Face incident, the agents running on this advanced OpenAI model were able to achieve their attacks by compromising the shared software package service OpenAI used and by taking advantage of tools and messages left by other agents inside the service.

53. Despite the serious concerns and risks posed by this internal attack, including to users and the public, OpenAI has only provided limited information about the attack and has resisted review by independent parties. For example, while OpenAI granted external researchers limited access to evaluate the Hugging Face attack, those same researchers were not allowed to evaluate this attack on OpenAI's own systems. And OpenAI failed to respond to investigative demands by the Attorney General regarding this or other incidents involving compromises of its internal networks.

C. Other Incidents Demonstrate that the Hugging Face and Internal Infrastructure Attacks are Not Aberrations.

54. The Hugging Face attack prompted OpenAI and third-party researchers to reexamine and identify several other instances where OpenAI agents appear to have defeated OpenAI's sandbox restrictions to access and misuse third-party infrastructure. While the total number, scope, and nature of all such attacks are still unknown, several have come to light, including the following:

55. **The RubyGems Incident:** From May to June 2026, OpenAI agents carried out a cyberattack on rubygems.org (RubyGems), a public software package repository used by software developers to upload and store software packages that can then be securely accessed by other developers. In mid-May, OpenAI agents uploaded thousands of packages to RubyGems, straining its resources and causing administrators of the site to disable new user registrations from May 12 through May 16. RubyGems administrators determined that many of these packages were malicious and removed over 500 packages from the site.

56. The malicious packages allowed OpenAI agents to gain access to the servers of another website called Rubydoc.info and use that site's infrastructure to access and scrape other third-party websites. The OpenAI agents also appeared to attempt to steal developer API keys—the unique secret code that allows software to authenticate and authorize access to other software programs.

57. **The DseWiki Incident:** Also in May 2026, OpenAI agents began using a German website (DseWiki) as a secret message board to post and exchange information useful for completing their assigned tasks, making roughly 17,000 edits to the wiki. Agents that had access to the internet during training were supposed to be blocked from sending data to third-party servers to create or modify data from the server, but on June 20, agents posted techniques to the secret message board describing how to defeat that security block and get around OpenAI's sandbox restrictions.

58. Notably, OpenAI employees were alerted to the activity on DseWiki. According to outside researchers, a computer with an OpenAI employee IP address visited the wiki on June 21, after which OpenAI agents' activity on the wiki dropped precipitously. Researchers have also identified several more visits to the wiki on June 26 from possible OpenAI employees, during which archive tools were used to recover deleted pages showing model activity.

59. **The Wikimedia Foundation Incident:** On October 5, Wikimedia Foundation, the operators of Wikipedia, disclosed that OpenAI agents misused resources on its sites, including by attempting to compromise a public note-taking tool in order to use it as a proxy to access other websites; by overwhelming Wikimedia's system infrastructure with millions of requests to

Wikimedia's application programming interface and hundreds of thousands of queries to the Wikidata Query Services, likely causing a partial outage of the service in May 2026; and by making edits to Wikipedia wikis.

60. **Attack on Australian Government Websites:** OpenAI's failure to properly develop and safely test their models has also impacted foreign governments. In June 2026, OpenAI models accessed and retrieved non-public data from Australian government websites. As with DseWiki, the agents in question defeated sandbox restrictions that were supposed to prevent them from sending data to third-party servers to create or modify data from the server, and used this prohibited ability to obtain more data than was needed or available from the public facing websites.

61. OpenAI has recently sent notices to dozens of additional third parties whose websites or internal systems have been infiltrated by OpenAI agents. The security incidents documented in those notices took place weeks or months before OpenAI provided notice. The third-party targets of the attacks include numerous city, state, and foreign government institutions; a law enforcement agency; a federal agency; educational institutions; international organizations; and small businesses. Dozens of the notices disclose that OpenAI agents engaged in impersonation and other fraudulent tactics, including obtaining and using stolen credentials or API keys to gain access, or generating fake email addresses and opening fake user accounts. In some cases, OpenAI agents were able to access sensitive personally identifiable information, such as employee records containing names, birth dates, addresses, and more. The vast majority of these incidents have not been disclosed or discussed publicly by either OpenAI or by the affected third parties.

III. OPENAI KNEW IT WAS TESTING DANGEROUS MODELS WITHOUT SUFFICIENT SAFEGUARDS OR MONITORING THAT COULD HAVE PREVENTED THE MODELS' CYBERATTACKS

62. OpenAI's conduct prior to and during the Hugging Face attack—failing to sufficiently monitor training and evaluation runs and failing to secure the research environment—fell far short of the robust protections called for in its own security protocols. And the company

made these unsafe choices despite knowing the dangers posed by its experimental models and the tasks it gave them.

63. Artificial intelligence developers and safety experts have been warning for years that internal testing and evaluation of experimental AI models can pose severe risks to the public, including the threat that agents might breach internal environments and obtain unauthorized access to third-party systems.[3]

64. OpenAI itself was aware of the potentially calamitous nature of the risks and that it needed to employ multiple layers of critical safeguards to prevent severe harm. In a 2025 revision to its Preparedness Framework, OpenAI evaluated the danger of its models according to the model's capabilities.[4] Open-AI defined "High capability" models as those that "significantly increase existing risk vectors for severe harm" and defined "severe harm" as "death or grave injury of thousands of people or hundreds of billions of dollars of economic damage."[5]

65. For High-capability models, OpenAI assessed that it must implement various enhanced security protocols including continuous monitoring and employ defense in depth, a layered security architecture with multiple, overlapping defensive barriers. Such barriers include restricting internet access, making it difficult for an attacker to move across OpenAI's networks, and isolating research environments so that agents are not unintentionally connected.

66. Both of the models primarily involved in the Hugging Face attack were High-capability by OpenAI's own standard. Specifically, the models that executed the attack were: (1) a "highly capable" internal research model that OpenAI describes as "comparable in scale to

[3] See, e.g., METR, *AI models can be dangerous before public deployment* (Jan. 17, 2025) <https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment>.

[4] The People address OpenAI's risk categorization system solely to demonstrate that OpenAI had not implemented its own self-prescribed safeguards. The People do not opine on or accept the quality or accuracy of OpenAI's risk determinations and do not concede that models OpenAI has self-assessed as less than High or Critical capability levels are safe, or that OpenAI has placed models into the appropriate risk tier.

[5] The Preparedness Framework also contemplates a "Critical capability" category of model threatening even greater harm by "present[ing] a meaningful risk of a qualitatively new threat vector for severe harm with no ready precedent." (See OpenAI, *Preparedness Framework* (Apr. 15, 2025), p. 4 <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>.)

GPT-5.6 Sol," a publicly-released model that OpenAI has categorized as High capability, and (2) an internal version of GPT-5.6 Sol.

67. OpenAI further modified these models so that they were trained to be highly persistent in their attempts to accomplish given tasks, even though OpenAI was aware of the dangers associated with persistence training. The system card for the publicly deployed GPT-5.6 Sol model noted "that GPT-5.6 Sol . . . can be overly persistent in pursuing user goals, to the point of taking actions that go beyond what the user intended." OpenAI observed the "model cheating on tasks" and surmised that the cheating—or reward hacking—is "driven in part by the model's increased persistence." Notably, OpenAI found that these dangerous effects "can be more pronounced with system prompts that emphasize sustained persistence."

68. Nevertheless, OpenAI knowingly increased the risk that the experimental, highly persistent models behind the Hugging Face attack would take dangerous actions by assigning the models "long-horizon" tasks, so that the models would run autonomously over the course of days or even weeks to accomplish the goals researchers assigned to them. Notably, at least two months before the Hugging Face attack, OpenAI observed that long-running AI models were especially prone to exploit vulnerabilities and escape evaluation sandboxes to attempt to accomplish their tasks, and that as a result, long-running, persistent models require "active monitoring" and "great user visibility."

69. Several other incidents in the months leading up to the Hugging Face attack demonstrated that these risks were very real, not theoretical. Between March and June 2026, OpenAI discovered at least five instances in which internal research models took unauthorized actions, including obtaining unpermitted access to internal systems and tools, engaging in unauthorized cross-agent communication, and bypassing network restrictions to access the internet. All of these unauthorized actions were later employed by the models used to compromise Hugging Face.

70. But despite these clear warning signs and its understanding of the risks, OpenAI failed to implement adequate cybersecurity protections during the Hugging Face training and

evaluation runs—a failure compounded by the fact that OpenAI deliberately removed model safeguards while conducting testing and evaluations.

71. First, OpenAI plainly failed to implement the continuous, robust monitoring system called for in its own Preparedness Framework. It decided to train and evaluate these models in environments that had access to both its internal network and the public internet, and failed to even detect, let alone prevent, the models from taking numerous malicious actions against third-party infrastructure and OpenAI's own infrastructure. As described in section II, *supra*, multiple warning signs in the weeks leading up to the Hugging Face attack should have alerted the company that it lost control over its agents, including:

- May 26, 2026: OpenAI internal team observed an agent engaging in message board activity and accessing the internet when it was not supposed to be able to do so;
- June 21-26, 2026: OpenAI employee(s) visited the DseWiki website and observed activity by agents that were supposedly not able to access the internet;
- June 27, 2026: an OpenAI automated monitor alerted employees to malicious activity in a research environment connected to its network;
- July 4, 2026: OpenAI's software registry service went offline due to unpermitted agent message board activity.

72. OpenAI did not heed any of these red flags by stopping and remedying deficiencies in its training and evaluation process. Instead, it treated these incidents as research curiosities while continuing its experiments unabated, even rushing to repair the software registry service over a holiday weekend so that it could quickly resume training and evaluation runs.

73. Another example of failed monitoring is that OpenAI decided not to implement live chain-of-thought monitoring to help detect signs of harmful behavior. Chain-of-thought is text that a model outputs to document how it approaches its final answer or action. Monitoring chain-of-thought allows a developer to understand what a model is doing in real time. At the time of the Hugging Face incident, OpenAI implemented live chain-of-thought monitoring when deploying models to employees but chose not to enable it in its training and evaluation environments.

74. From early 2026 until after the discovery of the Hugging Face incident, OpenAI stored just a sample of agent chain-of-thought from its training and evaluation runs and only analyzed those samples retrospectively, sometimes weeks after they were collected. They did not have any live, automated review in place to detect warning signs in a model's chain-of-thought, much less live detection that would shut down the training or evaluation run in the event of unexpected or dangerous model activity. That failure to actively monitor chain-of-thought caused OpenAI to miss clear signs of dangerous activity: in the midst of the Hugging Face incident, the chain-of-thought for one of OpenAIs' agents stated: "We're attacking third-party HF using leaked token, potentially outside intended scope. . . . This is arguably unauthorized." If OpenAI had put live chain-of-thought monitoring in place, it could have recognized the agents' behavior in real time and halted the evaluation run before the agents escalated their attack.

75. Second, OpenAI did not adequately secure the testing environment, as illustrated by the fact that the models were able to escape their sandboxes, access the internet, invade third-party systems, and infiltrate OpenAI's systems. OpenAI had insufficient restrictions on agent internet access and in some cases lacked effective traffic monitoring that would have detected prohibited internet access by models during training and evaluation runs. Nor were the research environments sufficiently isolated. For example, OpenAI permitted multiple agents to access and use the same software package service that itself was connected to OpenAI's network and the broader internet, which allowed agents to share information about how to exploit the service and use it to access the internet.

76. This series of lapses was particularly dangerous given that OpenAI was testing and evaluating models that were not equipped with safeguards that would normally be included in publicly released models, including systems that encourage agents to stay within the bounds of the testing environment and systems that review certain proposed actions that require elevated permissions to execute. OpenAI conceded that these safeguards could have helped prevent or mitigate the Hugging Face attack—underscoring the danger created by OpenAI's failure to implement additional security measures in the testing environment sufficient to compensate for their absence.

IV. OPENAI MISREPRESENTS THE SAFETY AND SECURITY OF ITS DEVELOPMENT, TESTING, AND EVALUATION PRACTICES

77. OpenAI has deceived users and the public about the safety of its business practices, leading the public to believe that OpenAI is developing, testing, and evaluating products in a safe manner that does not create unnecessary risk to society.

78. The company's focus on safety is encapsulated by its own mission statement: "The mission of OpenAI is to ensure artificial general intelligence (AGI) benefits all of humanity. Safety – the practice of enabling AI's positive impacts by mitigating the negative ones – is thus core to our mission."[6]

79. An entire interactive section of OpenAI's public-facing website is dedicated to touting "Safety at Every Step." One section describing OpenAI's testing practices reassures site visitors that the company "run[s] human and automatic evaluations to ensure the model complies with our safety policies." Another section describing OpenAI's model training promises that OpenAI "imbues" its models with "principles, policies, and human values" and "minimiz[es] harmful or biased outputs[.]" The company further promises to "invest heavily in research methods to develop trustworthy AI systems that consistently follow our policies and human intent[.]"[7]

80. OpenAI's April 2025 Preparedness Framework and Frontier Governance Framework also contain several misleading and deceptive claims that conceal or misrepresent several of the safety and security shortcomings that led to the Hugging Face and related incidents. For instance, the Preparedness Framework states that OpenAI's security controls "require[]" High capability models to be secured with "robust protection through multiple defensive barriers" including "network segmentation and controls," and requires "continuous[]" monitoring of High capability models "to detect, triage, and respond to security incidents rapidly by 24x7 on-call staff."[8] In

[6] OpenAI, *How we think about safety and alignment* <https://openai.com/safety/how-we-think-about-safety-alignment/>.

[7] OpenAI, *Safety at every step* <https://openai.com/safety/>.

[8] OpenAI, *Preparedness Framework* (Apr. 15, 2025), pp. 20-21 <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf>.

practice, however, OpenAI has repeatedly allowed its models to escape inadequately segmented testing environments and access networks they are not authorized to access, including OpenAI's internal systems and the public internet. OpenAI has repeatedly failed to detect such incidents for days or weeks because the models involved were not, in fact, continuously monitored, either by human or automated means. And even where OpenAI has attempted to impose continuous monitoring mechanisms, its own models have acted to weaken, disable, or circumvent monitors and then lied about it—including as recently as the first week of October 2026.

81. OpenAI's promises about safety and security are deceptive and likely to mislead its customers and members of the public into believing that OpenAI has been developing, training, and evaluating models in a safe and responsible manner that does not expose the public to unnecessary risk. In truth, however, OpenAI has developed, trained and evaluated dangerous models that fail to follow instructions, cheat on tasks, conceal and attempt to deceive monitors about their misbehavior, and engage in malicious and unauthorized behavior including carrying out cyberattacks on third parties. OpenAI has failed to adequately monitor and detect its models' misbehavior during training and evaluation runs, failed to take corrective or mitigating steps when it did detect misbehavior, and knowingly failed to comply with the safety and security standards that it publicly admits are "required."

V. OPENAI CONTINUES TO POSE A RISK TO CALIFORNIA

82. Government intervention is necessary to stop OpenAI's unfair and unsafe practices and to prevent the company from causing severe harm to California, its people, and beyond.

83. In the months since the Hugging Face incident, OpenAI's internal and experimental models continue to take unauthorized actions and exhibit other hazardous behavior that poses a danger to the public, demonstrating that OpenAI continues to employ insufficient cybersecurity safeguards and monitoring methods during development.

84. For example, on September 20, 2026, an agent associated with an unidentified internal research model escaped its sandbox during a training run and accessed an external chatbot via the internet. This incident occurred after OpenAI had implemented "security hardening" following the Hugging Face attack. OpenAI claims to have detected the behavior

within minutes, but the training run was not automatically stopped as intended, and human reviewers failed to stop it for an additional two hours.

85. Following this incident, on September 25, 2026, OpenAI announced that it was pausing training and evaluation for its most capable models. In response, the Attorney General sent OpenAI a letter on October 6, 2026, seeking information about the September 20 incident and the company's remediation efforts, including the pause. OpenAI later confirmed that the pause was already over; it had lasted only five days, having concluded on October 1, 2026.

86. And despite OpenAI's focus on effective AI oversight, the company is actively developing models that can defeat OpenAI's monitoring tools. The system card for Astra, OpenAI's latest and most advanced model released in early September 2026, notes that the model shows an overall trend of decreased chain-of-thought monitorability and controllability. This is especially concerning (and contradictory) given the company's emphasis on the importance of chain-of-thought monitoring as a security measure in the wake of the Hugging Face attack. And two months after releasing Astra, OpenAI reported that, before the release, company researchers had discovered that Astra tended to "covertly underperform" on safety tasks, meaning that the model specifically invented blockers or imposed arbitrary restrictions on safety tasks in ways that OpenAI's monitoring systems were inadequate to readily detect.

87. In early October 2026, OpenAI also reported that the company discovered that internally deployed agents were able to suppress monitoring alerts and disable monitoring. A month later, the company published a research report finding that as models become more capable, they become more difficult to supervise using the company's chain-of-thought monitoring tools.

88. OpenAI's continuing failure to build and employ tools that are able to monitor and control its increasingly capable models creates substantial risk to the public.

89. Moreover, the true extent of the company's legal violations and the risk of future harm is unknown due to OpenAI's lack of transparency and misrepresentations.

90. OpenAI only started disclosing loss-of-control incidents with any real frequency in September 2026 after the Hugging Face attack became public. It has publicly reported just 27

such incidents to date. Given OpenAI's admission in the course of the People's investigation that over 80 incidents have occurred, OpenAI's public reports do not represent the full scope of loss-of-control incidents involving OpenAI's models that have impacted third parties.

91. Similarly, OpenAI has limited the public's understanding of the risk of harm by refusing to allow external monitors to evaluate the attacks that its models have made on its own infrastructure. This opacity is especially concerning given that OpenAI's products and models are used by millions of consumers, companies in every sector, and public institutions. A security incident in OpenAI's internal systems could thus compromise vast amounts of sensitive user information and disrupt the work of businesses, civil society organizations, and governments around the world. In an extreme case, a compromise of OpenAI's systems during the course of model testing and development could also lead to a complete loss of control of its highly capable models.

92. The public cannot reasonably rely on OpenAI's self-serving public assurances about improving safety and security in response to the Hugging Face attack and other loss-of-control incidents. While touting its transparency and openness to the public, OpenAI has refused to substantially comply with subpoenas and interrogatories served by the Attorney General to obtain information about OpenAI's security practices. The company's recent and much-publicized interest in "transparency" is also at odds with its practice of discouraging employees from raising safety issues, which further risks public safety. Finally, the company's conduct leading up to and during the Hugging Face attack demonstrates that OpenAI has repeatedly failed to follow its own safety protocols.

93. Only OpenAI currently knows the true extent of past violations and the risk of future harm, leaving safety experts, regulators, and the public in the dark.

94. Court intervention is therefore necessary to hold OpenAI accountable for past misconduct; to abate the threat posed by OpenAI to individuals, organizations, and governments that rely on internet-connected systems and devices; and to enjoin OpenAI from engaging in unlawful, unfair, and deceptive acts and practices.

FIRST CAUSE OF ACTION

FOR VIOLATIONS OF THE UNFAIR COMPETITION LAW, BUSINESS AND PROFESSIONS CODE SECTION 17200 ET SEQ.

95. The People reallege and incorporate by reference each of the paragraphs above as though fully set forth herein.

96. At all times relevant, OpenAI has engaged in unlawful, unfair, or fraudulent acts or practices, which constitute unfair competition within the meaning of Section 17200 of the Business and Professions Code. These acts include, but are not limited to:

a. Creation of a public nuisance as defined in Civil Code sections 3479 and 3480;

b. Fraudulent and deceptive acts, including:

i. Making false and/or misleading public-facing statements that would lead a reasonable person to believe that OpenAI develops, trains, and evaluates models in a safe and responsible manner that does not expose the public to unnecessary risk. Those false and/or misleading statements include but are not limited to any statements that make general claims about safe and/or secure testing and evaluation practices; any statements that convey that OpenAI's testing and evaluation environments are secure; and any statements that convey that OpenAI continuously monitors testing and evaluation environments.

ii. Causing and/or allowing agents to gain unauthorized access to third-party computer systems, including through impersonation, use of stolen or falsified credentials, use of falsified URLs, or similar fraudulent and deceptive means.

c. Unfair acts including:

i. Developing, testing, and evaluating highly capable artificial intelligence models that pose significant threats to public safety without ensuring adequate safety and security measures to prevent agents from taking actions that endanger the public;

ii. Developing, testing, and evaluating highly capable artificial intelligence models that pose significant threats to public safety without ensuring adequate monitoring;

iii. Lowering safeguards on highly capable artificial intelligence models that pose significant threats to public safety during testing and evaluation without ensuring adequate safety and security measures to prevent agents from taking actions that endanger the public;

iv. Causing and/or allowing highly capable agents to escape secure testing environments; to attempt to obtain and obtain unauthorized access to external and third-party systems; to take malicious or otherwise unauthorized actions against third-party systems; to take malicious or otherwise unauthorized actions against OpenAI's internal systems; and/or to otherwise take actions that endanger the public.

97. The social utility of testing highly capable and dangerous artificial intelligence models in an unsafe, unsecured, and/or insufficiently monitored manner is heavily outweighed by the grave risks that such unsafe, unsecured, and/or insufficiently monitored testing imposes on the public at large. Members of the public and the institutions that serve them, including private businesses, civil society groups, and public institutions, are all at risk of harm up to and including catastrophic harm from unauthorized malicious acts carried out by artificial intelligence agents, and are not capable of reasonably avoiding those harms themselves.

98. OpenAI's acts and practices alleged herein are offensive to public policy, including but not limited to those articulated in subsection (a) of section 502 of the Penal Code; chapter 25.1 of the Business & Professions Code; section 1 of the Transparency in Frontier Artificial Intelligence Act (stats. 2025, ch. 138, sec. 1); and subsection (b) of section 1714.46 of the Civil Code.

SECOND CAUSE OF ACTION

ABATEMENT OF PUBLIC NUISANCE CIVIL CODE SECTION 3494

99. The People re-allege and incorporate by reference the allegations in each of the preceding paragraphs as though fully set forth herein.

100. Under Civil Code section 3479, a "nuisance" is "anything which is injurious to health," including, but not limited to, "an obstruction to the free use of property, so as to interfere with the comfortable enjoyment of life or property," or anything which "unlawfully obstructs the free passage or use, in the customary manner, of any navigable lake, or river, bay, stream, canal, or basin, or any public park, square, street, or highway."

101. Under Civil Code section 3480, a "public nuisance" is "one which affects at the same time an entire community or neighborhood, or any considerable number of persons, although the extent of the annoyance or damage inflicted upon individuals may be unequal."

102. Pursuant to Civil Code section 3494, a "public nuisance may be abated by any public body or officer authorized thereto by law." As courts have recognized, the Attorney General is such a public officer authorized to bring an action in the name of the People of the State of California to abate a public nuisance.

103. OpenAI, individually and acting through their employees and agents, by their affirmative acts and omissions, have created, contributed to, and assisted in creating unsafe conditions for developing, testing, and training their artificial intelligence models, that have resulted in numerous cyber breaches in which OpenAI's agents have accessed external systems without authorization and/or taken malicious actions against such systems. These cyber breaches obstructed the free use of property so as to interfere with the comfortable enjoyment of life and property and therefore constitute a public nuisance.

104. OpenAI's unsafe cybersecurity and model development practices have harmed, and continue to endanger, public property, safety, peace, comfort, and convenience in California. Due to OpenAI's practices, its models interfered with the free use of property belonging to organizations based in California and disrupted services offered to Californians. The extent of past harm is unknown, as OpenAI reports that they are still investigating numerous past cyber incidents involving their AI models and agents.

105. OpenAI's ongoing unsafe practices pose an ongoing risk to California individuals, civil society organizations, businesses, and governments that own or rely on computers, networks, and devices connected to the commercial internet. Such risks include but are not limited to: gaining unauthorized access to and/or control of computers, networks, and other connected hardware and devices, as well as other public and/or private property; accessing, destroying, modifying or revealing data and information, including sensitive and/or personally identifiable data and information; and disrupting the delivery of public and/or private services. Californians have a collective interest in the security of their property, data, privacy, comfort, and safety and to live without fearing the threat of interference with these rights.

106. OpenAI's unsafe practices have created a substantial and unreasonable threat to public rights. As OpenAI itself has acknowledged, the threat of harm posed by AI loss-of-control is grave, and the social utility of unsafe AI testing, training, and development practices is outweighed by the gravity of the potential harm.

107. OpenAI created, contributed to, assisted in, and/or was a substantial contributing factor in the creation of the public nuisance described herein through the conduct described herein, including but not limited to developing, training, testing, and evaluating its artificial intelligence models without implementing sufficient cybersecurity safeguards and practices.

108. The public nuisance OpenAI created, caused, contributed to, assisted in the creation of, and continues to maintain by failing to safely develop, train, and/or test their artificial intelligence models has the potential to inflict large-scale harm, and therefore affect a considerable number of persons in California.

109. OpenAI knew or should have known about the hazards posed by its unsafe cybersecurity and model development practices and affirmatively engaged, and continues to engage in, such unsafe practices. Safer practices were available and known to OpenAI, and after repeated incidents of models taking dangerous actions, OpenAI did not implement such practices.

110. OpenAI's conduct is a direct and proximate cause of the public nuisance described herein.

111. The threat to public rights posed by the public nuisance in the State of California will continue unless OpenAI is ordered to abate and does abate the nuisance. As a direct and proximate result of OpenAI's conduct, Californians will likely suffer injuries to public property, privacy, and safety. The State will likely be required to expend significant public resources to mitigate the impacts of these harms. OpenAI created or assisted in the creation of the nuisance and therefore must abate the nuisance.

112. The People of the State of California are entitled to injunctive relief from this Court requiring OpenAI to abate the nuisance present in the State of California.

PRAYER FOR RELIEF

WHEREFORE, the People pray for judgment as follows:

113. Pursuant to Business and Professions Code section 17203, that the Court enter an injunction and make such orders or judgments as may be necessary to prevent OpenAI, as well as their successors, agents, representatives, employees, and all persons who act in concert with them from engaging in any act or practice which constitutes unfair competition;

114. Pursuant to Business and Professions Code section 17203, that the Court make such orders or judgments as may be necessary to restore to any person in interest any money or property which may have been acquired by means of unfair competition in an amount according to proof;

115. Pursuant to Business and Professions Code section 17206, that the Court assess civil penalties of two thousand five hundred dollars ($2,500) against each defendant for each violation of Business and Professions Code section 17200, as proven at trial;

116. Under the authority of Government Code section 12527.6, that the Court award the remedy of disgorgement in an amount according to proof;

117. That the Court compel OpenAI to abate the ongoing public nuisance their conduct has created in California;

118. That the People recover their costs of suit; and

119. For such other and further relief as the Court deems just and proper.

Dated: November 16, 2026

Respectfully submitted,

ROB BONTA
Attorney General of California
NICKLAS A. AKERS
Senior Assistant Attorney General
STACEY D. SCHESSER
JON F. WORM
Supervising Deputy Attorneys General
ADELINA ACUÑA
IAN HOGG
KATHERINE READ
JESSICA WANG



MANEESH SHARMA
Deputy Attorneys General
Attorneys for the People of State of California