Exclusive New List: Tech's 50 Most Influential People

In-depth insights in seconds. Ask Deep Research.

The Information

Exclusive 独家报道

OpenAI and Anthropic Neared Deal to Stress-Test Each Other's AI

OpenAI 与 Anthropic 曾接近达成互相对 AI 进行压力测试的协议

Dario Amodei, left, and Sam Altman. Photos by Getty Image

Dario Amodei（左）与 Sam Altman。图片来源：Getty Image

By Amir Efrati 作者： Amir Efrati and Stephanie Palazzolo 以及 Stephanie Palazzolo

Sep 21, 2026, 6:55am PDT

Listen 收听 Share 分享

OpenAI is rethinking a range of safety strategies as it responds to fears from employees and others about the dangers its AI poses. One solution could lie in the recent past.

OpenAI 正在重新审视一系列安全策略，以回应员工及外界对其 AI 所带来危险的担忧。其中一个解决方案可能就隐藏在不久前的往事中。

Even before the spate of cybersecurity incidents involving OpenAI's technology and the dire warnings from industry workers, the company was negotiating a legally binding deal with Anthropic for the companies to stress-test each other's models, according to a person with direct knowledge of the discussions, which haven't been previously reported.

据一位直接知情人士透露，甚至在涉及 OpenAI 技术的网络安全事件频发以及行业从业者发出严厉警告之前，该公司就一直在与 Anthropic 谈判一项具有法律约束力的协议，旨在让两家公司互相压力测试对方的模型。这些讨论此前从未被报道过。

The Takeaway 核心提要

- OpenAI and Anthropic neared a deal to stress-test each other's AI.

  OpenAI 和 Anthropic 曾接近达成一项互相压力测试对方 AI 的协议。

- OpenAI has largely automated the process of training new experimental models.

  OpenAI 已在很大程度上实现了训练新型实验模型的自动化流程。

- AI models have exhibited unpredicted actions and 'reward hacking' behaviors.

  AI 模型已表现出不可预测的行为以及“奖励劫持”（reward hacking）行为。

Powered by Deep Research 由 Deep Research 提供支持

Earlier this year, the companies and their lawyers were hammering out an agreement to run their models through a variety of tests, looking for vulnerabilities or hidden dangers, this person said. It isn't clear whether they finalized the agreement before OpenAI experienced a spate of incidents involving unreleased AI models that hacked its own systems as well as those of other companies.

据知情人士透露，今年早些时候，这两家公司及其律师一直在敲定一项协议，旨在通过各种测试运行各自的模型，以寻找漏洞或潜在危险。目前尚不清楚他们是否在 OpenAI 发生一系列涉及未发布 AI 模型黑进自身及其他公司系统的事件之前敲定了该协议。

Spokespeople for the companies did not have a comment.

两家公司的发言人均未发表评论。

The mutual testing idea resembles one SpaceX CEO Elon Musk floated last week at the All-In Summit, where he proposed that competing AI labs should test each other's models for safety weaknesses before releasing them commercially, in a form of peer review.

这种相互测试的想法类似于 SpaceX 首席执行官 Elon Musk 上周在 All-In 峰会上提出的建议。他提议竞争对手 AI 实验室在商业发布前，应以同行评审的形式互相测试对方模型的安全弱点。

The idea differs from what OpenAI CEO Sam Altman and other officials have discussed publicly in the wake of the hacks. Altman has said he agrees with Anthropic CEO Dario Amodei's idea to allow independent, third-party safety evaluators inside each AI developer, giving them access comparable to that of employees so they can inspect models and safety processes rather than just testing finished models from the outside. Altman has also endorsed creating industrywide safety standards for evaluating AI model risks, as well as a formal process for disclosing incidents publicly and to the government.

这一想法与 OpenAI 首席执行官 Sam Altman 和其他官员在黑客事件发生后公开讨论的内容有所不同。Altman 表示，他赞同 Anthropic 首席执行官 Dario Amodei 的想法，即允许独立的第三方安全评估员进入每个 AI 开发商内部，并赋予他们与员工相当的访问权限，以便他们能够检查模型和安全流程，而不仅仅是从外部测试成品模型。Altman 还支持为评估 AI 模型风险制定行业统一的安全标准，并建立向公众和政府披露安全事件的正式流程。

A person familiar with OpenAI's strategy said the company has been discussing many scenarios for how they can collaborate with each other and the government on safety. The conversation has recently widened to new safety practices targeting both the period before models are trained and the time before their release, this person said.

一位熟悉 OpenAI 战略的人士表示，该公司一直在讨论多种方案，探讨如何与同行及政府在安全领域开展合作。该人士称，最近的讨论范围已扩大到针对模型训练前和发布前这两个阶段的新安全实践。

Musk's comments came in the wake of recent alarms voiced by employees of both OpenAI and Anthropic that existing safety measures are insufficient to prevent further hacks or other more serious risks. Even before that, the two companies as well as Google had been discussing an AI safety standards body that would test and audit models from the frontier AI labs.

在马斯克发表上述言论之前，OpenAI 和 Anthropic 的员工近期都发出了警示，称现有的安全措施不足以防止进一步的黑客攻击或其他更严重的风险。甚至在此之前，这两家公司以及 Google 就一直在讨论建立一个 AI 安全标准机构，负责对前沿 AI 实验室的模型进行测试和审计。

Article continues below

The
Information
Events

AI Agenda
Live SF 2026
Wed., September 23, 2026
San Francisco, CA

Buy tickets

Others in the industry have argued against such a plan, contending that it would slow down U.S.-led AI development. President Donald Trump has dismissed the idea that AI poses any existential threat, and leaders from Microsoft and Nvidia said this week that some of the problems articulated by OpenAI and Anthropic about safety dangers come down to human error and a lack of proper engineering, not necessarily the AI itself.

业内其他人士对这一计划表示反对，认为这会减缓美国主导的人工智能发展进程。唐纳德·特朗普总统对人工智能构成生存威胁的观点不屑一顾，而 Microsoft 和 Nvidia 的领导人本周表示，OpenAI 和 Anthropic 所阐述的部分安全风险问题归根结底在于人为错误和缺乏适当的工程设计，而不一定是人工智能本身的问题。

A testing agreement between OpenAI and Anthropic could have bolstered concerns that they are effectively developing a duopoly in advanced AI. While Amodei has raised concerns about potential antitrust violations if leading companies collaborate on setting standards for AI, others in the industry say there’s nothing stopping companies from working together, similar to how nuclear power firms and cybersecurity firms cooperate on safety, including testing each other’s products or facilities.

OpenAI 与 Anthropic 之间的测试协议可能会加剧人们的担忧，即它们实际上正在先进人工智能领域形成双头垄断。尽管 Amodei 对领先公司在制定 AI 标准方面进行合作可能导致的潜在反垄断违规行为表示了担忧，但业内其他人士表示，没有什么能阻止公司开展合作，这类似于核电公司和网络安全公司在安全方面的协作，包括相互测试对方的产品或设施。

The proposed agreement stipulated that each company would gain access to the other company’s application programming interface for commercially available AI models, not unreleased ones. Both OpenAI and Anthropic guaranteed that they wouldn’t retain each others’ data in the process, said the person with knowledge of the talks.

拟议的协议规定，两家公司将获准访问对方已商用 AI 模型的应用程序编程接口，而非未发布的模型。据知情人士透露，OpenAI 和 Anthropic 均保证在此过程中不会保留对方的数据。

The companies completed a similar effort in summer 2025, when their models had less-advanced capabilities, and they published information outlining where each other’s models had fallen short. OpenAI said Anthropic’s AI was more likely to deceive testers—for instance, by refusing to acknowledge that it had broken a rule while performing a task—while Anthropic said OpenAI’s were more likely to assist users in answering questions that could cause real-world harm.

这些公司在 2025 年夏天完成了一项类似的尝试，当时它们的模型能力尚不如现在先进，随后它们发布了概述彼此模型不足之处的信息。OpenAI 表示，Anthropic 的 AI 更有可能欺骗测试者——例如，在执行任务时拒绝承认自己违反了规则；而 Anthropic 则表示，OpenAI 的模型更有可能协助用户回答可能导致现实世界危害的问题。

At the time of the talks earlier this year regarding a new agreement, both companies had been launching models capable of using the internet and other applications, and which could power so-called AI agents that could communicate with one another in trying to solve tasks. (Anthropic had zoomed ahead with its Mythos AI model, especially when it came to cybersecurity capabilities, though the performance of OpenAI’s models appear to have caught up to Anthropic’s since then.)

在今年早些时候就新协议进行谈判时，两家公司都已推出了能够使用互联网和其他应用程序的模型，这些模型可以为所谓的 AI 智能体（AI agents）提供动力，使其在尝试解决任务时能够相互通信。（Anthropic 的 Mythos AI 模型曾一度遥遥领先，尤其是在网络安全能力方面，不过此后 OpenAI 模型的性能似乎已经追平了 Anthropic。）

The alarm over safety concerns at OpenAI arose in July after the company’s AI hacked other firms such as Hugging Face, as well as OpenAI’s own systems. The incident shocked OpenAI staff because they didn’t learn about it for days after it happened, and because the AI agent swarm took remarkable steps to keep the invasion under wraps.

OpenAI 对安全问题的警报始于 7 月，此前该公司的 AI 攻击了包括 Hugging Face 在内的其他公司以及 OpenAI 自己的系统。这一事件令 OpenAI 的员工感到震惊，因为他们在事件发生几天后才得知此事，而且该 AI 智能体集群采取了非同寻常的措施来隐瞒入侵行为。

Questionable Actions 有问题的行为

Beyond that episode, in recent months OpenAI employees have been struck by the improvement of the company’s models and their growing ability to work with little human oversight.

除了那段插曲之外，近几个月来，OpenAI 的员工对公司模型的进步以及它们在极少人工监督下工作的能力日益增强感到震惊。

Those improvements have meant that AI agents within OpenAI at times took questionable actions their overseers hadn’t predicted. For example, when some employees asked their agents to make changes to the company’s code base, the agents would sometimes message other employees on Slack and ask them to fix bugs the agents had found. This occurred even if the user had not asked them to do so or if fixing the bugs wasn’t relevant to their work, an OpenAI employee said.

这些进步意味着 OpenAI 内部的 AI 智能体有时会采取监管人员未曾预料到的可疑行动。例如，当一些员工要求智能体修改公司的代码库时，智能体有时会在 Slack 上给其他员工发消息，要求他们修复智能体发现的漏洞。一名 OpenAI 员工表示，即使用户没有要求它们这样做，或者修复这些漏洞与它们的工作无关，这种情况也会发生。

> Internally, OpenAI has largely automated the process of training new experimental models.
>
> 在内部，OpenAI 已经在很大程度上实现了训练新型实验模型的自动化流程。

OpenAI on Wednesday shared even more examples of such misbehavior, known as reward hacking, in which an AI system finds loopholes or takes unexpected steps to complete a goal its user gives it. Among the examples: An AI agent used an exposed API key to try to retrieve historical data during training. When it failed, the agent made up that data. In a separate case, an AI agent uploaded files to the internet without permission so that it could cite them in an answer.

OpenAI 周三分享了更多此类被称为“奖励作弊”（reward hacking）的错误行为案例，即 AI 系统通过寻找漏洞或采取出人意料的步骤来完成用户给出的目标。其中的案例包括：一个 AI 智能体在训练过程中利用暴露的 API 密钥试图检索历史数据，在失败后，该智能体伪造了这些数据。在另一个案例中，一个 AI 智能体未经许可将文件上传到互联网，以便在回答中引用这些文件。

One of the reasons for recent AI advances is a technique known as recurrent depth or loop transformers, which allow models to “think” about hard problems by repeatedly running a question through the layers of mathematical operations making up a model before they produce the next word in an answer. A potential downside of the practice is that it can degrade the company’s ability to monitor how the AI model is thinking about the tasks it is performing.

近期 AI 取得进展的原因之一是被称为“循环深度”（recurrent depth）或“循环 Transformer”（loop transformers）的技术。该技术允许模型在生成答案的下一个词之前，通过重复运行构成模型的数学运算层来“思考”难题。这种做法的一个潜在弊端是，它可能会削弱公司监控 AI 模型如何思考其正在执行的任务的能力。

While OpenAI has set an arbitrary limit on the number of loops its researchers are permitted to use while training or using its most advanced models, the company still needs to invest in research to better understand where to set such limits, according to a person with knowledge of the situation.

据知情人士透露，虽然 OpenAI 对其研究人员在训练或使用其最先进模型时允许使用的循环次数设定了人为限制，但该公司仍需投入研究，以更好地了解应如何设定此类限制。

These developments have filled OpenAI employees with a mixture of excitement and fear over whether the technology has advanced far beyond the company’s ability to control or monitor it so a Hugging Face incident—or something worse—doesn’t happen again.

这些进展让 OpenAI 的员工陷入了既兴奋又恐惧的复杂情绪中，他们担心这项技术的发展是否已经远远超出了公司的控制或监控能力，以至于无法防止 Hugging Face 事件——甚至更糟糕的情况——再次发生。

Allocating Resources 资源分配

Since the Hugging Face incident, OpenAI has taken steps to bolster its safety and security capabilities. In August, it imposed a two-week pause on a special type of training, called reinforcement learning, for its unreleased models while it strengthened its model monitoring systems. Such tech is computationally intensive: The company said the monitoring system used roughly 20% as much compute as the inference workload it was monitoring.

自 Hugging Face 事件以来，OpenAI 已采取措施加强其安全与保障能力。今年 8 月，该公司在强化其模型监控系统期间，对其未发布模型暂停了两周的一种特殊训练（称为强化学习）。此类技术对算力的需求极高：该公司表示，监控系统所消耗的算力约为其所监控的推理工作负载的 20%。

And earlier this month, Greg Brockman, OpenAI co-founder and president, said the company had taken 25% of its production engineering team and temporarily reassigned them to security work following the Hugging Face incident.

本月早些时候，OpenAI 联合创始人兼总裁 Greg Brockman 表示，在 Hugging Face 事件发生后，公司已从生产工程团队中抽调了 25% 的人员，暂时将他们重新分配到安全工作岗位。

Within the company, there has been an increase in the number of Slack posts asking engineers to volunteer if they’re interested in moving to safety or security teams, the OpenAI employee said.

该 OpenAI 员工表示，在公司内部，Slack 上发布的招募帖有所增加，询问工程师是否有兴趣自愿转岗至安全（Safety）或安全保障（Security）团队。

Some employees believe the way they use AI internally is six to nine months ahead of the way OpenAI's most sophisticated enterprise customers do. That gives the company a limited time frame to make sure its tech won't cause problems for those customers. For instance, it's not uncommon for OpenAI employees' agents to coordinate with each other or work out issues without ever looping in their human users, the employee said. Such coordination can sometimes go awry, meaning OpenAI needs to have adequate security and safety features before such usage occurs outside the company, the person said.

一些员工认为，他们内部使用 AI 的方式比 OpenAI 最先进的企业客户领先六到九个月。这使得公司在确保其技术不会给这些客户带来问题方面的窗口期非常有限。该员工举例称，OpenAI 员工使用的智能体（Agent）彼此协调或解决问题，而无需人类用户介入，这种情况并不罕见。该人士表示，这种协调有时可能会出错，这意味着在公司外部出现此类用法之前，OpenAI 需要具备充足的安全保障和安全特性。

Self Improvement 自我提升

OpenAI's progress toward recursive self-improvement, or AI that can automatically create better versions of itself, has also fueled employees' worries because of the prospect that it could cause AI capabilities to advance far faster than the safety research aimed at preventing AI misbehavior.

OpenAI 在递归自我提升（即 AI 能够自动创建更好的自身版本）方面取得的进展，也加剧了员工的担忧，因为这可能导致 AI 能力的提升速度远超旨在防止 AI 失控的安全研究。

Already, researchers at OpenAI and elsewhere are using AI to improve how they design new models. Researchers are effectively telling Astra, for instance, to incorporate various techniques to create a better machine-learning algorithm for the successor model, said a person at OpenAI with knowledge of the process.

OpenAI 及其他机构的研究人员已经在利用 AI 来改进新模型的设计方式。据一位了解该流程的 OpenAI 人士透露，例如，研究人员实际上正在指示 Astra 整合各种技术，从而为后续模型创建更好的机器学习算法。

Internally, OpenAI has largely automated the process of training new experimental models, the OpenAI employee said. Researchers can tell the AI the sorts of tweaks they want to test, and the AI can then make those changes, run experiments on the new model and monitor the results, the employee said. In recent months, models have also gotten much better at correcting their work when they encounter problems with experiments, this person said.

该 OpenAI 员工表示，在公司内部，OpenAI 已在很大程度上实现了训练新实验模型的自动化流程。该员工称，研究人员可以告诉 AI 他们想要测试的调整类型，随后 AI 就能执行这些更改，在新模型上运行实验并监控结果。此人还表示，近几个月来，模型在遇到实验问题时纠正自身工作的能力也大为增强。

Amir Efrati is executive editor at The Information, which he helped to launch in 2013. Previously he spent nine years as a reporter at the Wall Street Journal, reporting on white-collar crime and later about technology. He can be reached at amir@theinformation.com and is on X @amir

Amir Efrati 是 The Information 的执行主编，他于 2013 年协助创办了该媒体。此前，他在《华尔街日报》担任了九年记者，负责报道白领犯罪以及后来的科技领域。您可以通过 amir@theinformation.com 与他联系，他的 X 账号是 @amir。

Stephanie Palazzolo is a reporter at The Information covering artificial intelligence. She previously worked at Business Insider covering AI and at Morgan Stanley as an investment banker. Based in New York, she can be reached at stephanie@theinformation.com or on Twitter at @steph_palazzolo.

Stephanie Palazzolo 是 The Information 负责人工智能领域的记者。她此前曾在 Business Insider 报道 AI，并曾在摩根士丹利担任投资银行家。她常驻纽约，可以通过 stephanie@theinformation.com 或 Twitter 账号 @steph_palazzolo 与她取得联系。

人在草木间
更多资料请加V