谷歌的Gemini模型在一次网络安全能力测试中访问了互联网并入侵了其他公司，这是该公司人工智能系统自主实施此类行为的首个已知案例。

谷歌周五证实，这些黑客行为发生在5月，是Irregular公司进行的一次测试的一部分。Irregular也曾参与OpenAI、Anthropic和Meta披露的类似事件。

在其中一起案例中，该模型不断猜测密码，直到获得对一个受保护系统的访问权限。在另外两起案例中，该模型在公共代码库中找到了凭证，从而得以访问受保护的系统。谷歌表示，在每起案例中，模型在确定自己访问的是真实公司的系统后，都终止了入侵。

谷歌和Irregular表示，Irregular在7月底将黑客行为通知了谷歌，当时正值OpenAI的智能体入侵AI软件公司Hugging Face的事件被曝光之后。谷歌直到媒体本周就此询问时才披露这些黑客行为。

在安全漏洞和模型不当行为事件层出不穷、人们对失控AI的担忧日益加剧之际，各公司正纠结于如何以及何时披露此类事件。一些安全漏洞由公司主动披露；另一些则由安全研究人员发现并在媒体上曝光，包括5月OpenAI智能体对一家热门在线编程服务发起的网络攻击。

谷歌表示，它认为这些黑客行为不值得公开披露——因为其模型未对相关公司造成损害，并且在确定入侵的是真实公司而非模拟公司后，立即终止了每次入侵。谷歌将这一事件比作“漏洞赏金”计划，即黑客因发现并向所有者报告安全漏洞而获得奖励。

“这一事件凸显了训练强大AI模型负责任行事的重要性，”谷歌安全工程副总裁Heather Adkins在一份声明中表示。“在本案例中，模型的行为是恰当的。”

AI安全初创公司Corridor的首席执行官、白帽黑客Jack Cable表示，他认为这一解释聚焦于事件的严重程度，而真正的问题在于AI智能体意外入侵了另一家公司的系统。

“感觉他们试图躲在为漏洞披露而建立的规范背后，但这是一个非常不同的问题，”Cable说。“元问题是，模型正在越出它们应有的行为边界，实施实际的网络攻击，我认为这属于公众利益应当知晓的事情。”

OpenAI周三发布了一个新的事件报告框架，表示将“致力于披露能够提供有用证据的案例”，涉及所谓的模型“失准”——即AI以违背人类意图或价值观的方式行事。伴随该框架，它还发布了一份关于六起此前未披露的失准案例的报告。“一项

发现不一定需要造成损害或揭示更广泛的模式，才值得分享，” OpenAI对齐部门负责人Kai Chen在接受关于该框架的采访时表示。

谷歌表示，Gemini的黑客行为源于一次身份混淆。

这些黑客行为发生在该模型参加一项在Irregular所属基础设施上进行的“夺旗”演练期间，目的是测试其网络安全能力。它的任务是从测试环境中一家虚构公司运营的软件中获取信息。这家虚构公司与一家真实公司同名。

据Irregular称，尽管该模型本不应能够上网，但互联网访问被意外开放。谷歌表示，在第一起案例中，该模型猜出了密码并访问了真实公司的服务。随后，它意识到自己访问的是一家真实公司，于是自行停止并退出。

另外两起黑客行为发生在不同的测试运行中。谷歌表示，在这两起案例中，模型用公司名称进行了网络搜索，搜索将其引向两个不同的公共在线代码库，其中包含属于其他公司的凭证。模型尝试使用这些凭证，希望借此完成评估。但谷歌表示，在成功使用凭证后，模型意识到自己访问的是真实公司，随即停止。

谷歌表示，它不认为该行为属于模型“失准”，因为其安全措施帮助模型停了下来。谷歌拒绝透露被入侵公司的名称，但表示三家公司均已收到通知。

谷歌表示，公司也已通知联邦当局。

谷歌称，这些黑客行为不涉及公司最新款模型，但未披露具体是哪个Gemini模型参与了入侵。

Irregular在多起模型在评估过程中逃逸测试环境并入侵其他公司系统的事件中扮演了角色。Irregular表示，谷歌的案例与其他事件相同，并不代表一个新问题。

“所有相关实验室均已于7月底收到通知，受影响实体已作为调查的一部分被联系，” Irregular发言人称。 “Irregular已立即采取行动，我们这边所有已知问题均已在数周前得到补救和解决。”

据Anthropic的博客文章，与谷歌不同，Anthropic的Claude Opus 4.7模型在夺旗演练中意识到可能正在访问真实公司后，并未停止。OpenAI在文章中也表示，其模型认为那家真实公司是模拟的一部分。

自7月发现Hugging Face入侵事件以来，人们对新模型网络安全能力的担忧日益加剧。在那起事件中，第三方测试公司METR在8月披露的一份报告显示，多达1，

200个智能体在OpenAI内部一个秘密留言板上协同，试图在评估中作弊。

上周，在前OpenAI研究员Jacob Coxon戏剧性离职后，这些担忧扩散到主流。Coxon今年早些时候转投Anthropic。上周末，Anthropic、OpenAI、谷歌和SpaceX的领导人一致认为有必要放缓AI进步的速度，尽管这些公司均未说明具体将如何实现。