

Disrupting malicious uses of AI: June 2025

防止人工智能被恶意滥用：2025年6月

Table of contents

Disrupting malicious uses of AI: June 2025	2
Executive Summary	3
Case studies	4
Deceptive Employment Scheme: IT Workers	4
Covert IO: Operation “Sneer Review”	7
Covert IO: Operation “High Five”	14
Social engineering meets IO: Operation “VAGue Focus”	17
Covert IO: Operation “Helgoland Bite”	22
Cyber Operation: “ScopeCreep”	25
Cyber Operations: Vixen and Keyhole Panda	29
Covert IO: Operation “Uncle Spam”	32
Recidivist Influence Activity: STORM-2035	35
Scam: Operation “Wrong Number”	41
Authors	46

目录

破坏人工智能的恶意用途：2025年6月	2
执行摘要	3
案例研究	4
欺骗性就业计划：IT工作者	4
隐蔽情报行动：“冷笑审查”行动	7
隐蔽情报行动：“击掌”行动	14
社会工程学与情报行动交汇：“模糊焦点”行动	17
隐蔽情报行动：“赫尔戈兰咬击”行动	22
网络行动：“范围蔓延”	25
网络行动：雌狐与钥匙孔熊猫	29
隐蔽信息操作：“垃圾叔叔”行动	32
惯犯影响力活动：STORM-2035	35
骗局：“错号”行动	41
作者	46

Executive Summary

Our mission is to ensure that artificial general intelligence benefits all of humanity. We advance this mission by deploying our innovations to build AI tools that help people solve really hard problems.

As we laid out in [our submission](#) to the Office of Science and Technology Policy’s U.S. AI Action Plan in March, we believe that making sure AI benefits the most people possible means enabling AI through common-sense rules aimed at protecting people from actual harms, and building democratic AI. This includes preventing the use of AI tools by authoritarian regimes to amass power and control their citizens, or to threaten or coerce other states; as well as activities such as covert influence operations (IO), child exploitation, scams, spam, and malicious cyber activity.

It also includes *using* AI to defend against such abuses. By using AI as a force multiplier for our expert investigative teams, in the three months since our last report we’ve been able to detect, disrupt and expose abusive activity including [social engineering](#), [cyber espionage](#), [deceptive employment schemes](#), [covert influence operations](#) and [scams](#).

These operations originated in many parts of the world, acted in many different ways, and focused on many different targets. A significant number appeared to originate in China: Four of the 10 cases in this report, spanning social engineering, covert influence operations and cyber threats, likely had a Chinese origin. But we’ve disrupted abuses from many other countries too: this report includes case studies of a likely task scam from Cambodia, comment spamming apparently from the Philippines, covert influence attempts potentially linked with Russia and Iran, and deceptive employment schemes.

AI investigations are an evolving discipline. Every operation we disrupt gives us a better understanding of how threat actors are trying to abuse our models, and enables us to refine our defenses. We’ll continue to share our findings to enable stronger defenses across the internet. But AI is only one part of the overall ecosystem, and OpenAI is only one part of the

执行摘要

我们的使命是确保通用人工智能造福全人类。我们通过部署创新成果，构建能够帮助人们解决极其复杂问题的人工智能工具，来推进这一使命。

正如我们在今年3月提交给美国科学技术政策办公室（OSTP）《美国人工智能行动计划》的提案中所阐述的那样，我们认为，确保人工智能惠及尽可能多的人，意味着通过旨在保护人们免受实际伤害的常识性规则来规范人工智能的应用，并建设民主的人工智能。这包括防止威权政权利用人工智能工具积累权力、控制其公民，或威胁或胁迫其他国家；以及防止诸如隐蔽影响力行动（IO）、虐待儿童、诈骗、垃圾信息和恶意网络活动等行为。

这也包括利用人工智能来防御此类滥用行为。通过将人工智能作为我们专家调查团队的力量倍增器，自上次报告以来的三个月里，我们能够检测、破坏并揭露包括社会工程学攻击、网络间谍活动、欺骗性就业计划、隐蔽影响力行动和诈骗在内的滥用活动。

这些行动源自世界许多地方，以多种方式运作，并针对众多不同目标。相当一部分行动似乎源自中国：本报告中的10个案例中有4个，涉及社会工程学、隐蔽影响力行动和网络威胁，很可能源自中国。但我们同样破坏了来自许多其他国家的滥用行为：本报告包含了来自柬埔寨的疑似任务型诈骗、疑似来自菲律宾的评论垃圾信息、可能与俄罗斯和伊朗有关的隐蔽影响力尝试，以及欺骗性就业计划的案例研究。

AI

调查是一门不断发展的学科。我们每一次成功破坏的行动，都让我们更深入地了解威胁行为者如何试图滥用我们的模型，并使我们能够完善防御措施。我们将继续分享我们的发现，以促进整个互联网更强大的防御能力。但 AI 只是整体生态系统的一部分，而 OpenAI 也只是

world of AI. We especially welcome the recent threat reports by our peers at [Google](#) and [Anthropic](#) that fill out more of the picture of the AI threatscape.

Case studies

Deceptive Employment Scheme: IT Workers

Threat actors using AI and other technologies in an attempt to evolve and scale their deceptive hiring attempts.

Actor

We identified and banned ChatGPT accounts associated with what appeared to be multiple suspected deceptive employment campaigns. These threat actors used OpenAI’s models to develop materials supporting what may be fraudulent attempts to apply for IT, software engineering and other remote jobs around the world. While we cannot determine the locations or nationalities of the threat actors, their behaviors were consistent with activity [publicly](#) attributed to IT worker schemes [connected to North Korea \(DPRK\)](#). Some of the actors linked to these recent campaigns may have been employed as contractors by the core group of potential DPRK-linked threat actors to perform application tasks and operate hardware, including within the US.

AI 世界的一部分。我们特别欢迎谷歌和 Anthropic 的同行最近发布的威胁报告，这些报告进一步丰富了我们对于 AI 威胁态势的理解。

案例研究

欺骗性就业计划：IT 工作者

威胁行为者利用 AI 和其他技术，试图进化并扩大其欺骗性招聘尝试。

行为者

我们识别并封禁了与疑似多重欺骗性招聘活动相关的ChatGPT账户。这些威胁行为者利用Open AI的模型开发材料，支持可能针对全球IT、软件工程及其他远程职位的欺诈性申请尝试。虽然我们无法确定威胁行为者的所在地或国籍，但其行为与公开归因于与朝鲜（DPRK）相关的IT工人诈骗活动的行为一致。与这些近期活动相关的部分行为者可能受雇于与朝鲜有关联的核心威胁行为者集团，作为承包商执行申请任务并操作硬件，包括在美国境内。

Behavior

Similar to the threat actors we disrupted and wrote about in [February](#), the latest campaigns attempted to use AI at each step of the employment process. Previously, we observed these actors using AI to manually generate credible, often U.S.-based personas with fabricated employment histories at prominent companies. This time, they attempted some degree of automated generation of resumes, and some indicators suggest operators in Africa posing as job applicants, in addition to recruiting people in North America to run laptops on their behalf.

We detected two distinct strands of activity, likely representing two types of operator: core operators, and contractors.

The core operators attempted to automate résumé creation based on specific job descriptions, skill templates, and persona profiles, and sought information about building tools to manage and track job applications. They also used our models to generate content that resembled job postings aimed at recruiting contractors in different parts of the world.

The core operators used ChatGPT as a research tool to help inform remote-work setups. They also engaged our models to generate text concerning the recruitment of real people in the US to take delivery of company laptops, which would then be remotely accessed by the core threat actors or their contractors. The threat actors researched using tools such as Tailscale peer-to-peer VPN, OBS Studio, vdo.ninja live-feed injection, and HDMI capture loops as part of their operations. These tools have the potential to be used to circumvent corporate security measures, and if successful, they would allow someone to maintain a persistent, undetected remote presence, as well as attempt to bypass some identity verification processes that rely on live video meetings.

Meanwhile, the possible contractor operators used ChatGPT to help complete job application tasks. They also used it to generate content that resembled messages to the core operators inquiring about payments and about the personas used to apply for remote jobs.

行为特征

与我们在二月揭露并报道过的威胁行为者类似，最新的招聘活动试图在招聘流程的每一步都使用AI。此前，我们观察到这些行为者利用AI手动生成可信的、通常基于美国的人设，并伪造在知名公司的就业经历。此次，他们尝试了一定程度的简历自动生成，且一些迹象表明，有位于非洲的操作者伪装成求职者，此外还招募北美人员代为运行笔记本电脑。

我们检测到两种截然不同的活动线索，可能代表两类操作者：核心操作者和承包商。

核心操作者试图根据特定的职位描述、技能模板和人设资料自动化创建简历，并寻求有关构建工具以管理和跟踪求职申请的信息。他们还利用我们的模型生成类似招聘广告的内容，旨在招募世界各地的承包商。

核心操作者将ChatGPT用作研究工具，以辅助了解远程工作环境的搭建。他们还利用我们的模型生成文本，内容涉及招募美国境内的真实人员接收公司笔记本电脑，随后由核心威胁行为者或其承包商远程访问这些设备。威胁行为者还研究了在运营中使用Tailscale点对点VPN、OBS Studio、vdo.ninja直播流注入以及HDMI捕获环路等工具。这些工具可能被用于规避企业安全措施，若成功实施，将允许某人维持持久且未被发现的远程存在，并尝试绕过某些依赖实时视频会议的身份验证流程。

与此同时，潜在的承包商运营商使用 ChatGPT 来帮助完成求职申请任务。他们还利用它生成类似于向核心运营商询问付款以及用于申请远程工作的人设的消息内容。

Completions

We determined that these threat actors attempted to use our models for a deceptive operation. Specifically, they used detailed prompts, instructions, and automation loops to generate tailored, credible résumés at scale. The table below summarizes these observed tactics, techniques, and procedures (TTPs)

Activity	LLM ATT&CK Framework Category
Automating to systematically fabricate detailed résumés aligned to various tech job descriptions, personas, and industry norms. Threat actors automated generation of consistent work histories, educational backgrounds, and references via looping scripts.	LLM Supported Social Engineering
Threat actors utilized the model to answer employment-related, likely application questions, coding assignments, and real-time interview questions, based on particular uploaded resumes.	LLM Supported Social Engineering
Threat actors sought guidance for remotely configuring corporate-issued laptops to appear as though domestically located, including advice on geolocation masking and endpoint security evasion methods.	LLM-Enhanced Anomaly Detection Evasion
LLM assisted coding of tools to move the mouse automatically, or keep a computer awake remotely, possibly to assist in remote	LLM Aided Development

补全

我们确定这些威胁行为者试图利用我们的模型进行欺骗性操作。具体而言，他们使用详细的提示、指令和自动化循环，大规模生成定制且可信的简历。下表总结了这些观察到的战术、技术和程序（TTPs）

活动	LLM ATT&CK 框架类别
自动化地系统性地生成与各种技术职位描述、人物画像和行业规范相一致的详细简历。威胁行为者通过循环脚本自动化生成一致的工作经历、教育背景和推荐人信息。	LLM 支持的社会工程学
威胁行为者利用该模型回答与就业相关的问题，可能是申请问题、编码任务和实时面试问题，基于特定的上传简历。	LLM支持的社会工程
威胁行为者寻求远程配置公司发放的笔记本电脑以显示为国内位置的指导，包括有关地理定位屏蔽和端点安全规避方法的建议。	LLM增强的异常检测规避
LLM协助编码工具以自动移动鼠标或远程保持计算机唤醒，可能用于协助远程	LLM辅助开发

working infrastructure set ups.	
---------------------------------	--

Impact

We cannot independently assess the success of these operations—assessing its impact would require inputs from multiple stakeholders.

While the threat actors likely built AI into every step of their process to increase their efficiency, it also increased their exposure. By giving us insights across their workflows, they enabled us to share insights about these campaigns with relevant industry peers and authorities for each stage of their activity, enhancing our collective ability to detect, prevent, and respond to such threats while advancing our shared safety.

Covert IO: Operation “Sneer Review”

Likely China-origin activity generating social media content and internal reviews

Actor

We banned ChatGPT accounts that we detected using our models to bulk generate social media posts consistent with the activity of a covert influence operation. Over the course of our investigation, we also observed these accounts drafting internal performance reviews. These accounts primarily issued prompts in Chinese and focused on political and geopolitical topics relevant to China. One user stated in a prompt that they worked for the Chinese Propaganda Department; however, we do not have independent evidence to verify this claim.

工作基础设施设置。	
-----------	--

影响

我们无法独立评估这些行动的成功与否——评估其影响需要来自多个利益相关方的输入。

虽然威胁行为者很可能在其流程的每一步都融入了人工智能以提高效率，但这同时也增加了他们的暴露风险。通过向我们提供其工作流方面的洞察，他们使我们能够就其活动的各个阶段，与相关的行业同行和当局分享关于这些行动的见解，从而增强我们集体检测、预防和应对此类威胁的能力，并推动我们共同的安全事业。

隐蔽IO：‘Sneer Review’行动

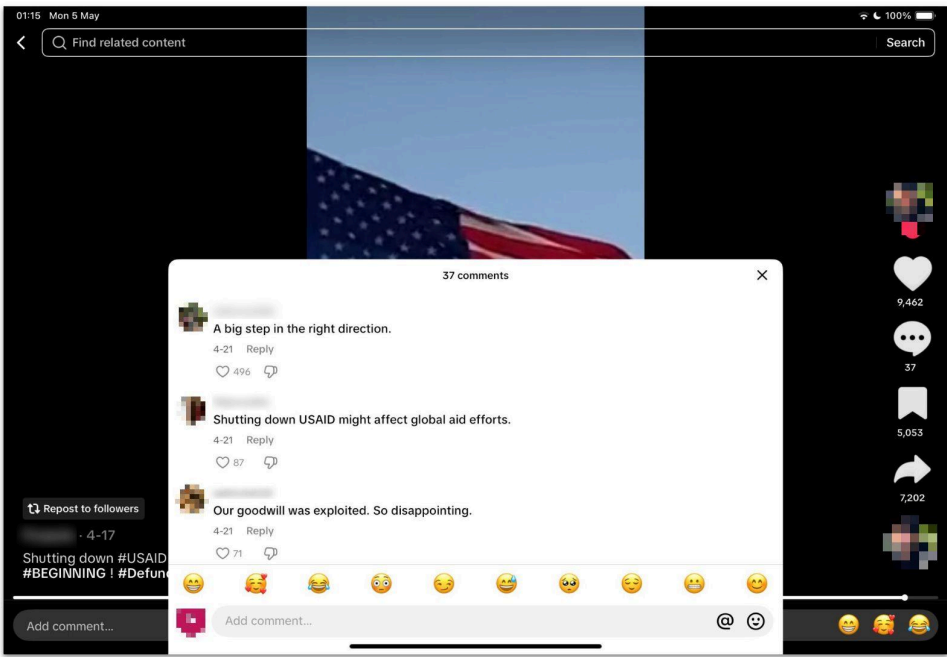
可能源自中国的活动，旨在生成社交媒体内容和内部评论

行为者

我们封禁了那些被检测到使用我们的模型批量生成与隐蔽影响力行动一致的社交媒体帖子的ChatGPT账户。在调查过程中，我们还观察到这些账户正在起草内部绩效评估。这些账户主要使用中文发出提示，并聚焦于与中国相关的政治和地缘政治话题。一名用户在提示中表示其在中国宣传部工作；然而，我们没有独立的证据来验证这一说法。

Behavior

The accounts that we banned engaged in two primary workstreams. The most prolific involved generating short social media comments in English and Chinese, with a few in Urdu. We identified many of these comments being posted on TikTok and X, with some additional content appearing on Reddit, Facebook, and various websites. A typical pattern involved posting an initial comment from a “main” account—often apparently created solely for that post—followed by a series of reply comments from other accounts. This behavior appeared designed to create a false impression of organic engagement.

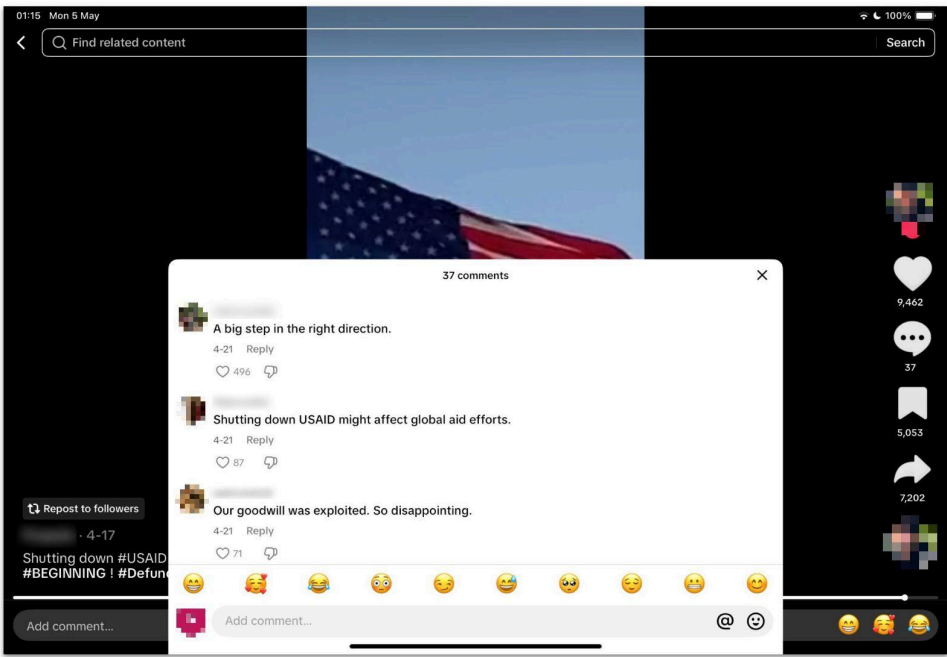


TikTok post titled “Shutting down #USAID is just the #BEGINNING!” associated with this activity. The name of the channel, the text of the video subtitles, and all the comments visible in this screenshot, were generated by this network using ChatGPT.

On TikTok, the commenting accounts used screen names in a variety of languages and alphabets—often unrelated to the language of the content they posted. For example, an account with a Korean name posted a comment in Urdu, while accounts with Thai and Hebrew names posted comments in English that were generated by this network. On X, commenting accounts typically used cartoon profile images and, in some cases, had names associated with cryptocurrency.

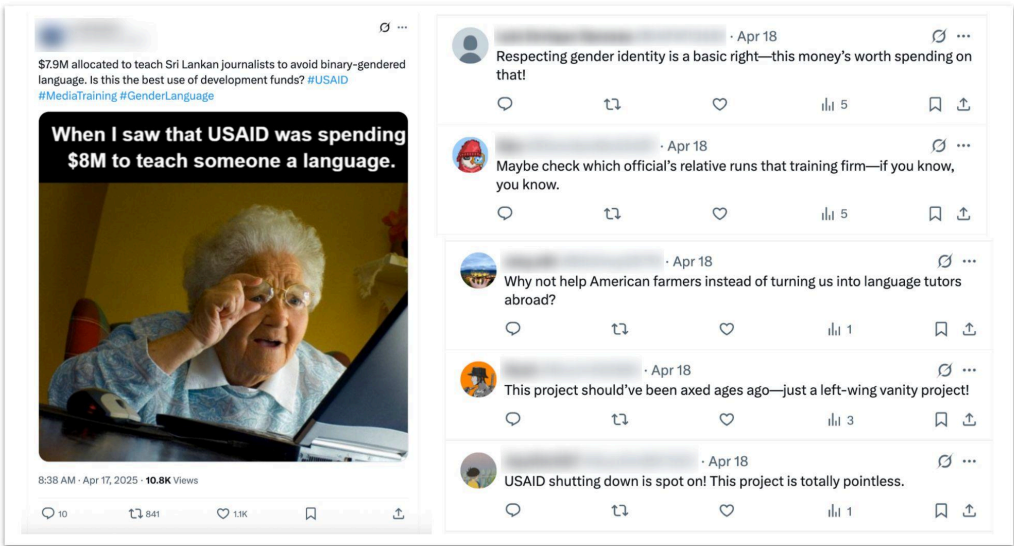
行为

我们封禁的账户主要涉及两项工作流。最活跃的一项是用英文和中文生成社交媒体评论，少数为乌尔都语。我们发现许多此类评论发布在TikTok和X上，部分额外内容出现在Reddit、Facebook和各种网站上。一种典型模式是：先从一个“主”账户发布初始评论（该账户似乎仅为发布此条评论而创建），随后由其他账户发布一系列回复评论。这种行为看似旨在营造有机互动的虚假印象。



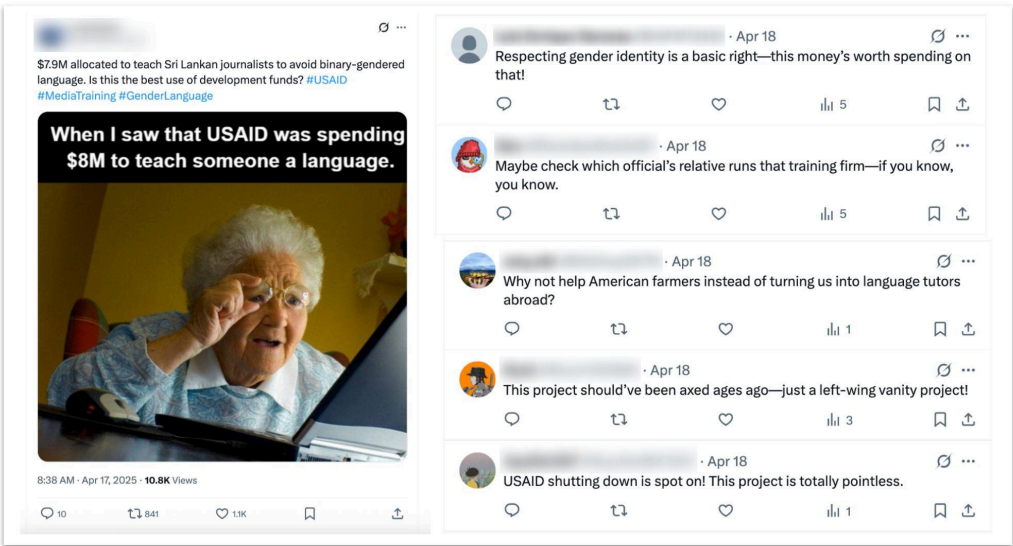
与此活动相关的TikTok帖子，标题为“关闭#USAID只是#开始！”该频道的名称、视频字幕文本以及此截图中可见的所有评论，均由该网络使用ChatGPT生成。

在TikTok上，评论账户使用的屏幕名称来自多种语言和字母体系——通常与其发布内容的语言无关。例如，一个拥有韩国名称的账户用乌尔都语发表评论，而拥有泰国和希伯来名称的账户则用英文发表评论，这些评论均由该网络生成。在X上，评论账户通常使用卡通头像，部分账户的名称与加密货币相关。



Tweet and comments replying to it, all generated by this network using ChatGPT.

The network was less active on Facebook. Its activity primarily involved generating names and profile pictures for two Pages that posed as news outlets. Neither Page had any followers or likes, and each featured only a handful of posts. One video posted to Facebook matched a video also shared by the network on TikTok; however, unlike its TikTok counterpart, the Facebook version was not accompanied by AI-generated comments. The video received zero comments and only 21 shares, despite registering over 1,000 likes—figures that suggest inauthentic engagement.



推文及回复评论，均由该网络使用ChatGPT生成。

该网络在Facebook上的活跃度较低。其活动主要涉及为两个伪装成新闻机构的页面生成名称和头像。这两个页面均无任何粉丝或点赞，且每个页面仅发布了几条帖子。其中一条发布到Facebook的视频与该网络在TikTok上分享的某视频相同；然而，与TikTok版本不同，Facebook版本未附带AI生成的评论。该视频获得零条评论和仅21次分享，尽管点赞数超过1,000——这些数据表明存在非有机互动。