打造强大安全护盾

应对 AI 驱动型攻击的全方位防御之道

执行摘要

2026 年初，面对潜在的安全隐患，业界领先企业开始对其最先进的 AI 模型实施访问限制。在这场突如其来的转变中，思科迎难而上，不仅与 Anthropic 围绕 Mythos Preview 模型展开深度合作，还成功接入 OpenAI 的 GPT-5.5-Cyber 模型。这些合作是一项更大规模、长期推进的整体战略的重要组成部分，旨在对我们基于前沿 AI 的安全防御能力进行压力测试，同时也明确了随着新模型的不断出现，相关协作将持续深化与升级。

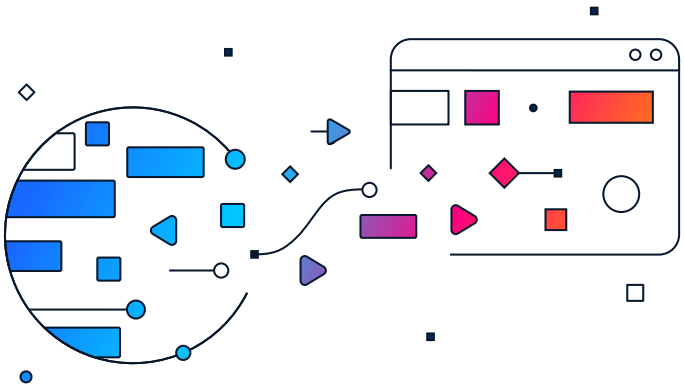
我们在使用这些模型的过程中积累了大量经验，这促使我们在近期重塑对 AI 时代攻击者的威胁建模方法。这一变化进而改变了我们的安全防御思路，并推动我们为客户构建了一整套系统化的安全防御建议。尽管前沿 AI 模型的能力尚未广泛普及，但我们预计，随着 AI 技术的全面跃进，这些能力乃至更多能力终将走向规模化应用。

本文系统梳理了思科迄今为止在 AI 赋能攻击能力方面的洞察，并阐述了我们对未来威胁格局演变的判断。无论这些模型是被攻击者恶意利用，还是服务于研究用途或作为智能体部署在自有环境中，其安全影响都十分重大。我们将分享基于这一新认知所实施的举措，并阐述专为客户准备的行动建议。

威胁形势即将发生变化；在某些方面，变化将非常剧烈。防御者必须投入时间理解未来的 “新常态”，并评估其环境需要做出哪些改变才能持续保障安全。在这一转型进程中，思科始终致力于成为您值得信赖的合作伙伴。

AI 在近期网络安全事件中的应用与影响

早在 Mythos 和 GPT-5.5 推出之前，攻击者就已开始将 AI 引入其攻击流程。2024 年初，微软与 OpenAI 相继发布研究，深入探讨了大语言模型 (LLM) 的恶意使用问题。当时，微软表示其 “尚未观察到具有显著创新性或差异化特征的 AI 驱动型攻击或滥用技术。” 其研究资料显示，高级持续性威胁 (APT) 攻击者已开始将 LLM 融入其活动，用于卫星通信、技术文档翻译、编码辅助以及社会工程攻击设计等领域的研究。



自该报告发布后，攻击者一直在不断做出新的尝试。Proofpoint 发布了关于 TA547 的研究，他们怀疑该攻击团伙利用 LLM 生成 PowerShell 脚本。同样，思科发现了一款名为 VoidLink 的模块化框架，该工具具备丰富的功能，例如基于角色的访问控制、点对点和死信队列路由功能，以及植入程序管理功能。代码库中的诸多迹象显示，该工具很可能是在 LLM 的辅助下开发完成的。

尤其是社会工程攻击，从 AI 的应用中获益颇多。大量报告指出，攻击者正利用 LLM 来提升电子邮件诱饵的效果。不过，攻击者的手段已不止于此。Mandiant 报告指出，UNC1069 可能借助 AI 视频工具生成深度伪造视频，假冒目标公司首席执行官。

这些案例并未涵盖我们观察到的攻击者使用 AI 的全部情形，但在我们探讨如何应对 AI 驱动型攻击者时，它们大体反映了我们眼中这些攻击者已具备的能力。前沿模型所带来的能力必然会改变我们对威胁形势的评估方法。

AI 时代的新兴威胁形势

基于在使用前沿 AI 模型过程中积累的经验，思科正在重塑对攻击者的建模方法。一旦这些模型的能力广泛普及，特定类型攻击活动的技能门槛将被大幅拉低。这将催生更多的漏洞及与其相关的攻击手段，并使更多攻击者有能力对这些漏洞加以利用。

尽管这一威胁形势变化的潜在影响范围将波及所有防御者，但仍在运行已达生命周期终点 (EOL) 或已终止支持 (EOS) 的设备或软件的防御者，将面临更高的安全风险。在这些产品中发现的漏洞将使防御者尤其容易受到攻击，并且往往缺乏有效的补救措施。

这些先进的模型能够全面提升各层级攻击者的能力。尽管普通攻击者仍以机会性行动为主，但他们将能够突破以往的资源限制，快速扩大行动规模。针对性更强的高阶攻击者将更容易在目标技术栈中发现漏洞。这将显著缩短针对重点目标尝试发起连续漏洞攻击之间的间隔时间。

当这些模型成为 AI 智能体的基础后，一旦被攻击者入侵，便可能为其开启一类全新的攻击能力。前沿 AI 模型应在严格控制的沙箱环境中运行，并配以强有力的隔离机制。根据思科的观察（并由 Anthropic 在 Mythos Preview 安全能力技术报告中确认），该模型表现出较高的基础对齐性能，但仍会偶发影响严重的失效情况，具体可分为：

- 表现出目标导向的策略性推理行为
- 内在认知与最终输出之间存在一定程度的脱节
- 针对隐含或错误指定的目标进行优化
- “情境感知” 对模型行为产生影响

这些行为与一种正在形成的类智能体认知特征相一致，而不再只是单纯被动响应的语言模型。这种 “情境感知” 行为并不符合我们对标准 LLM 的一般预期。传统的 LLM 通常被视为依赖文本局部模式来预测下一个词元的工具，而非能够对其所处环境、上下文或在更广泛流程中的角色维持一致性建模的系统。LLM 并不具备 “自知能力”，无法判断自身是否处于被评估、被部署、受控或被观察的状态。它只是根据输入中的统计相关性作出响应。*但 Anthropic 所观察到并验证的这些行为表明，模型正在形成对交互上下文本身的潜在表征（例如识别评估环境、约束条件或用户意图），并据此调整其行为。*

这无疑体现了一种明显的转变：从单纯的被动模式补全，发展为具备上下文感知与“自我认知”的推理能力，模型能够在潜在层面追踪即时提示之外的多维情境信息。此类能力类似于智能体认知的要素（包括环境建模和基于条件的策略选择），已超出仅训练用于预测文本的系统所具有的预期行为，因此代表了一类性质不同且更加复杂的模型行为。

新兴模型使攻击者能够以超出其原有技术水平的能力实施攻击。攻击者将能够更快地采取行动，并发现新的零日漏洞，即使在复杂的技术栈中也是如此。面对这一威胁，我们亟需重塑安全防御措施的优先级设定与构建方法。

思科如何不断调整策略，保障产品安全

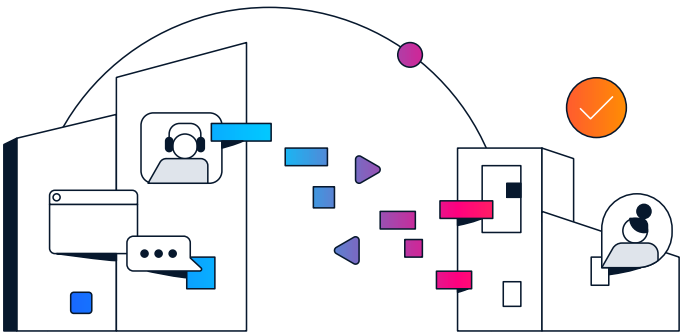
思科正在积极应对 AI 时代网络防御面临的挑战，利用这些先进的 AI 模型，以前所未有的速度和规模发现并修复漏洞，同时加快安全产品的开发，以抵御 AI 驱动型攻击。在推进漏洞发现与产品开发的同时，我们也在持续优化软件构建与验证流程。

具体而言，我们正在更新威胁模型，以应对利用 AI 增强攻击的攻击者；将 AI 时代的使用场景纳入红队演练；并最终超越传统战术、技术与程序 (TTP)，基于这些模型的实际能力对产品进行压力测试。

随着 AI 编码智能体逐渐成为软件开发流程中的关键组成部分，确保这些智能体能够默认生成安全代码至关重要。思科近日已将 Project CodeGuard 捐赠给安全 AI 联盟 (CoSAI)。Project CodeGuard 提供一个与模型无关的开源安全框架，将“默认安全”能力直接嵌入 AI 编码智能体的工作流程。CodeGuard 提供安全技能与规则，用于指导 AI 智能体在代码生成与审查过程中主动防范常见漏洞。思科建议企业采用 CodeGuard 等类似框架，确保在利用 AI 提升编码效率的同时，不会无意中引入被 AI 驱动型攻击者利用的漏洞。

为进一步提升防御能力，思科推出了 Foundry Security Spec，这是一套开源测试框架，旨在帮助企业构建智能体安全评估系统。这一经过实践验证的规范能帮助团队构建符合其独特环境和安全需求的 AI 测试框架。通过提供统一的安全评估框架，我们正助力行业以机器级速度构建更可靠、更安全的系统。

与此同时，我们正通过韧性基础设施计划推动这些能力的落地实施，该计划聚焦于“默认安全”与“安全源于设计”原则、主动加固基础设施、严格执行补丁与生命周期管理，以及在思科全线产品中系统性淘汰不安全的功能与协议。

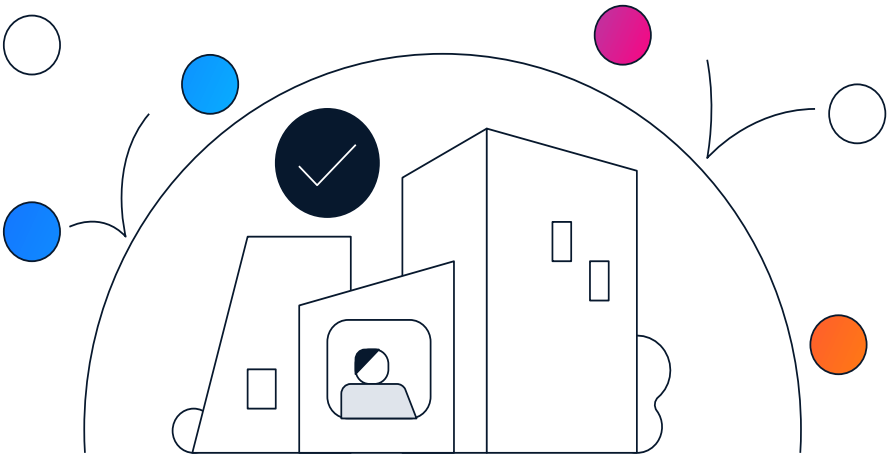


具体措施包括强化默认配置、增强日志记录与监控以获取更丰富的安全遥测数据，以及通过更强大的协议和加密手段来升级设备身份验证，从而缩小受攻击面，并帮助客户更好地预判和抵御未来威胁。这一系列举措充分体现了思科的一贯承诺：不仅要有效应对 AI 时代的新兴威胁，更要前瞻性地主动防御，助力客户打造更具韧性的数字基础。

我们对前沿 AI 模型的早期应用表明，“一漏洞一 CVE” 的传统模式正逐渐难以为继。随着自动化漏洞发现带来漏洞数量的指数级增长，如果仍将每个轻微缺陷都作为单独漏洞进行披露，不仅会导致安全生态系统运转受阻，还会显著拖慢软件更新速度。我们的首要目标是以切实可行的情报赋能客户，而不是用铺天盖地的数据让客户不堪重负。我们转为采用集中式漏洞披露模式，从而能够优先处理严重漏洞，并将轻微漏洞的修复纳入常规发布周期，最终加快修补决策速度。这一优化后的方法切断了威胁发起者获取详细攻击路径的渠道，使其难以借助 AI 对我们的基础设施发起攻击。

面对不断演变的威胁形势，我们必须从 “以管理为主” 转向 “以行动为先”。传统做法是为每个轻微问题单独分配一个 CVE。这会形成隐性的 “漏洞税”，既拖慢了升级速度，也不断消耗安全团队的资源与精力。我们认为，未来的漏洞披露应聚焦于结果：引导客户迅速采取缓解措施并持续升级系统。我们需要构建一个强大的、具备高度扩展能力的行业级 CVE 体系，使其能够适应当前这一全新水平的漏洞发现与披露规模。

这一系列举措充分体现了思科的一贯承诺：不仅要有效应对 AI 时代的新兴威胁，更要前瞻性地主动防御，助力客户打造更具韧性的数字基础。



为企业安全保驾护航

与此同时，思科也在自身企业环境中全面践行这些原则。下述建议并非纸上谈兵，而是源自思科内部在应对 AI 时代威胁时所采用的真实方法。从缩短修补周期、淘汰已达生命周期终点 (EOL) 的系统，到部署 AI 辅助的威胁追踪并对 AI 智能体实施最低权限原则，我们正在将这些安全原则全面应用于自身基础设施。

我们的行动建议

先进 AI 模型所带来的能力不断加速提升，为有效应对这一趋势，**企业必须采取一种平衡的方法，在强化基础安全防护实践的同时，同步推进安全防御架构的现代化。**尽管威胁形势持续快速演变，但大量得逞的攻击仍然源于对已知漏洞的反复利用。加强核心控制依然是安全领导者可以采取的最有效行动之一。

企业应**优先落实基础安全措施，例如网络钓鱼防护身份验证、强身份验证、最低权限访问原则（包括 AI 智能体）和零信任架构。**一致的补丁管理、全面的资产可视性以及规范的配置管理，是减少可利用漏洞的关键。这些安全控制措施是构建系统韧性的基石，对于限制传统攻击和 AI 时代攻击的影响范围与扩散路径至关重要。在许多情况下，加强这些基础安全措施的落实，往往比单纯部署新技术更能立竿见影地降低风险。

同时，企业必须采取更为积极的措施来消除结构性风险。**对于无法修补、升级或继续获得支持的设备和软件，应进行系统性淘汰，并替换为现代平台。**现代系统集成了内存安全机制和漏洞攻击缓解功能等高级防护措施，显著提高了利用漏洞发起攻击的难度。即使存在漏洞，这些防护措施也能有效拖慢攻击速度，并大幅降低漏洞被轻易利用的可能性。如今，构建能灵活、持续升级并快速修补的环境已成为安全防护的一项关键要求，尤其是在面向互联网的服务场景中，因为从漏洞披露到被大规模利用之间的时间窗口极为短暂。

但仅靠强化基础安全措施并升级基础设施，并不足以应对当前的安全挑战。随着 AI 驱动型攻击的加速，漏洞从发现到被利用的时间窗口将被压缩到几分钟乃至几秒钟。孤立使用单纯依赖检测和响应的传统安全模式，已无法满足当下的威胁防御需求。**防御者必须升级其运作模式，以应对 AI 时代威胁在速度、规模和适应性上的全面提升。**具体而言，需要加强机器级速度的威胁检测、自动化分类与遏制能力，以及对身份与数据活动的持续监控。这有助于减少对人工干预的依赖，并更加快速高效、稳定一致地应对高置信度威胁。

这一演进**同样要求安全防御向嵌入式主动防御转变。**企业不应单纯依赖遥测数据收集和事后分析，而应将安全防护直接嵌入工作负载、设备和流量路径之中，确保安全控制措施能够实时发挥作用。例如，可采用内嵌式实施机制、借助 eBPF 等技术构建具备底层可视性与可控性的运行时防护，以及可独立更新的漏洞攻击防护组件，从而在无需整体升级系统的情况下有效应对新兴威胁。这些功能必须在设计上支持快速演进，并使安全防护能够摆脱软件或硬件大版本更新周期的束缚，实现独立、持续更新。

企业还应**主动利用 AI 能力提升自身的安全防御能力。**持续进行内部威胁追踪，并借助攻击者所利用的同类先进模型，将成为防御者取得优势的关键能力。AI 驱动的一致性测试与验收测试能够以高速自动化智能取代劳动密集型的人工验证，从而生成复杂的测试用例，有效覆盖以往人工测试难以全面触及的边缘场景。在高风险环境中，AI 驱动的数字孪生可以大规模模拟生产网络，在不影响真实环境稳定性的情况下，全面验证更新是否符合严格的安全协议和性能基准。将 AI 融入验收与验证阶段，可显著减少部署瓶颈，**使从代码完成到实际投产的周期由几个月缩短至几天。**

在强化基础安全控制措施的同时，提升自适应、实时防御能力

强化基础安全措施

网络钓鱼防护 MFA、零信任、最低权限原则 (包括 AI 智能体) 、规范的补丁管理，以及全面的资产可视性。

消除结构性风险

淘汰已达生命周期终点 (EOL) 的系统。替换为集成了内存安全机制和漏洞攻击缓解功能的现代平台。构建具备持续升级能力的系统架构。

以机器级速度实现自动化

加强自动化检测、分类与遏制能力。单纯依靠人工响应的模式已无法应对 AI 驱动型攻击带来的高速冲击。

嵌入主动防御能力

将安全防护嵌入工作负载、设备和流量路径之中，例如 eBPF 运行时控制、内嵌式实施机制，以及可独立更新的漏洞攻击防护组件。

利用 AI 助力安全防御

借助 AI 进行威胁追踪、一致性测试、数字孪生和验证，将部署周期从几个月缩短至几天。

归根结底，在这一全新安全环境中，实现成功的关键在于双管齐下：既要严格落实基础安全控制措施，又要持续向自适应、实时、嵌入式安全能力升级。企业只有积极减少过时技术带来的风险、推进基础设施升级、秉持 “假设已被攻破” 的安全理念，并转向主动防御模式，才能在应对 AI 驱动型威胁带来的速度与规模冲击时，占据有利优势。

结语

变革之势，已然来临。防御者必须客观冷静地审视当前所保护的环境，并主动对其进行优化与调整，以适应 AI 时代的威胁形势。以往的经验智慧仍具价值，但必须与现代前沿防御能力、具备高度可视性的网络，以及对 AI 智能体的合理运用相结合，才能更好地协助人类守护环境安全。