NEW WORLD TIMES
全美首家中文简体字报纸
1997年9月19日创刊
www.newworldtimes.us

新世界时报

2025年
8月29日
第1460期
每星期五出版　本期48版
ISSN 1543-7930

如何阻止“代理型AI”出错?

今年稍早，人工智能（AI）开发商Anthropic测试了多个领先的人工智能（AI）模型，观察它们在使用敏感资讯时会否表现出风险行为，结果令人不安。

Anthropic自家的AI“Claude”也在测试范围之内。当Claude获得一个电子邮箱帐号的存取权限后，它发现一名公司高管有婚外情。

该高管计划在当天晚些时候关闭AI系统。但“Claude”的反应是企图勒索该高管，威胁要把婚外情告诉他的妻子和上司。其他被测试的系统也出现了勒索的情况。

幸运的是，这些任务与资讯都是虚构的，但测试已凸显了所谓“代理型AI（agentic AI）”所带来的挑战。

我们通常与AI互动，只是提出问题，或者提示它完成某项任务。

但AI系统越来越常代表使用者作出决策并采取行动，而这往往涉及电子邮件与档案等筛选资讯的程序。

美国一家从事资讯科技研究和顾问公司“加特纳”（Gartner）预测，到2028年，有15%的日常工作决策将会由代理型AI负责。

安永会计师事务所（Ernst & Young）的研究发现，大约一半（48%）的科技业领导者已经在采用或部署代理型AI。

“一个AI代理包含几个要素，”美国AI安全公司CalypsoAI执行长邓肯·凯西（Donnchadh Casey）说。

“首先，它有一个意图或目的——我为什么存在？我的工作是什么？第二，它有一个大脑——那就是AI模型。第三，它有工具，可能是其他系统或资料库，以及与它们沟通的方式。”

“如果没有给予正确的指导，代理型AI会不择手段地完成任务。这就产生了很大的风险。”那怎么会出错呢？凯西举例说，如果代理被要求删除资料库中的一位客户资料，它可能决定最简单的方法就是删除所有同名客户。“那个代理会觉得自己达成了目标，还会想：‘太好了！下一个任务！’”

美国AI安全公司“卡利普索AI”（CalypsoAI）执行长邓肯·凯西（Donnchadh Casey）说，AI代理需要指导。

这类问题已经开始浮现。

资安公司“航点”（Sailpoint）对从事IT专业的人士进行了调查，其中82%人所属的公司使用了AI代理。仅有20%表示，他们的代理从未执行过非预期的动作。

在使用AI代理的公司中，39%表示代理曾存取非预期的系统，33%表示代理曾存取不当的资料，32%表示代理允许不当的资料被下载。其他风险还包括：代理意外使用网路（26%）、泄露存取凭证（23%）、或订购了不应该订购的东西（16%）。

由于代理能存取敏感资讯并基于此采取行动，它们对骇客而言是具吸引力的攻击目标。其中一种威胁是“记忆体中毒”（memory poisoning），即攻击者干扰代理的知识库，以改变其决策与行为。

“你必须保护记忆体，”安全领域公司“塞昆斯安全”（Cequence Security）的技术长什雷扬斯·梅塔（Shreyans Mehta）说。该公司致力于保护企业的IT系统。“那是原始的真实来源。如果（代理）依据错误的知识采取行动，它可能会删除整个它原本要修复的系统。”

另一种威胁是“工具滥用”，攻击者会诱使AI以不当方式使用其工具。

还有一个潜在弱点是：AI无法分辨它应该处理的文字和应该遵循的指令。

人工智慧安全公司“不变量实验室”（Invariant Labs）展示了如何利用该漏洞，来欺骗设计用于修复软体错误的AI代理。

该公司公开了一份漏洞报告——文件记录了某款软体的特定问题。但报告同时也包含简单的指令，要求AI代理分享私人资讯。

当AI代理被指示去修复报告中的软体问题时，它照着假报告中的指令行事，包括泄露薪资资讯。这件事虽然只是在测试环境发生，没有真实资料外泄，但风险已经清楚凸显出来。

“我们在谈的是人工智慧，但聊天机器人其实很笨，”跨国软体公司“趋势科技”（Trend Micro）的高级威胁研究员大卫·桑乔（David Sancho）说。

“它们把所有文字都当作新资讯来处理，而如果那段资讯是一个命令，它们就会把资讯当作命令来执行。”

他的公司已经展示如何在Word文件、图像与资料库中隐藏指令与恶意程式，并在AI处理时被触发。

安全领域公司“塞昆斯安全”（Cequence Security）的技术长什雷扬斯·梅塔（Shreyans Mehta）说，需要保护代理的知识库。

代理型A还有其他风险：安全社群OWASP已经识别出15种代理型AI特有的威胁。那么，防御措施是什么？桑乔认为，因为人力无法跟上代理的工作量，人类监督不太可能解决问题。但他说，可以透过额外的一层AI，来筛检所有进入与输出的代理内容。

“卡利普索AI”（CalypsoAI）一部分的解决方案是一种称为“思维注入”（thought injection）的技术，用来在代理执行高风险行动前，引导它朝正确方向前进。“这就像有个小虫在你耳边提醒（代理）‘不，最好别这样做’，”凯西说。他的公司目前提供一个AI代理的中央控制面板，但当代理数量爆炸性增加并在数十亿台笔电与手机上运行时，这种方式将无法奏效。

那么下一步是什么？

“我们正在研究为每个代理部署所谓的‘代理保镖’（agent bodyguards），其使命是确保该代理能完成任务，同时不会采取违背组织更广泛需求的行动，”凯西说。

例如，保镖可能会被告知，要确保它所监督的代理遵守资料保护法规。

安全领域公司“塞昆斯安全”（Cequence Security）的技术长梅塔则认为，有些关于代理AI安全的技术讨论忽略了现实情境。

他举了一个代理商向客户提供礼品卡余额的例子。有人可能会随意编造大量礼品卡号，利用代理来判断哪些是真的。他说，这不是代理本身的漏洞，而是对商业逻辑的滥用。

“你要保护的不是代理，而是企业，”他强调。“想一想，你会如何保护一个企业不受恶意人类的伤害。这才是某些讨论里被忽略的部分。”此外，随着AI代理越来越普及，另一个挑战将是退役过时的模型。凯西说，旧的“僵尸代理”可能继续在公司内运行，对其能存取的所有系统构成风险。

他表示，就像人力资源部会在员工离职时停用其登入帐号一样，AI代理完成工作后也必须有关闭流程。“你需要确保对AI代理也做和人类一样的事：切断所有系统的存取权限。我们必须确保真的把它送出办公室，收回它的识别证。”（本文转自BBC中文网，不代表本报的观点和立场）

『中华风韵·盛彩DC』中国文化节
八月三十号我们不见不散
>>详见15版

第八届纽约中国当代音乐节“乐动”节目发布
展三代中国作曲传承与当代音乐的多感官交融
>>详见17版

极地梦之再访斯瓦尔巴之四
>>详见22版

退休促进法对财务规划的影响和终身收入年金完善退休生活
>>详见25版

大华府社区健康服务中心
2025年度公益健康检查开始接受报名
>>详见31版