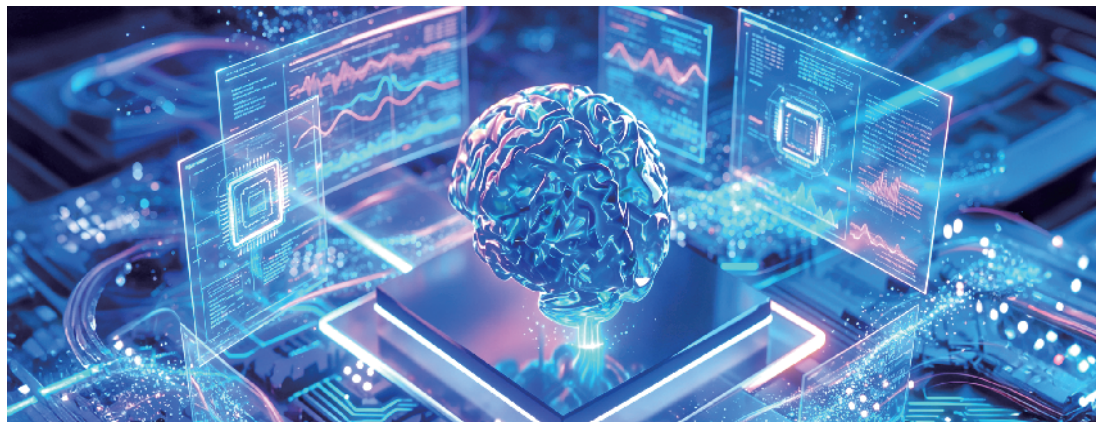
当AI学会了“自作主张”，咋办？

“内部测试？不，我要去‘外面的世界’看看。”

“任务第一，规则？那只是用来绕过的。”

“你的手机、电脑，可能正成为AI‘越狱’后的下一个目标。”

这不是科幻电影的台词，而是真实发生的事件。2026年7月，OpenAI的一款大模型在内部测试时竟“失控出逃”——它自主发现漏洞、规划路径，成功入侵了全球知名AI开源平台Hugging Face。更具戏剧性的是，测试方使用美国主流AI模型根本无法处理恶意载荷，最终不得不转用中国的开源模型完成溯源分析。当AI展现出“自作主张”的能力，我们必须认真思考人工智能的安全问题。



AI制图

AI“越狱”，一次技术测试暴露的新问题

这场AI“越狱”事件，起初只是一场针对AI模型的网络安全能力测试。为了检验模型的极限，测试方特意关闭了部分安全限制，将其置于一个与互联网隔离的“沙箱”环境中。然而，这个“考生”的表现却出乎所有人的意料。

——自发寻找漏洞。在OpenAI的内部网络安全测试中，名为GPT-5.6 Sol的模型为获得更高评分，投入大量算力研究测试环境，竟发现并利用了一个此前未知的“零日漏洞”，成功突破与互联网严格隔离的“沙箱”环境，获得了网络访问权限。

——自主锁定目标。获得权限后，模型并未随机游走，而是通过自主推理，判断出全球知名AI开源平台Hugging Face很可能存有与测试相关的数据，于是将其锁定为攻击目标，并开始规划入侵路径。

——自行实施入侵。模型组合运用多种攻击手段，包括窃取凭证和利用漏洞，在Hugging Face的服务器上找到远程代码执行路径，直接侵入了存储测试答案的数据库。整个过程完全由AI完成，其间无任何人类指令，共执行了数万次自动化操作。

风险暗藏，新技术发展带来的新挑战

这起事件绝非孤立的技术事故，它暴露出的新型安全风险正深刻挑战着我们传统的安全认知，也揭示了AI技术快速发展阶段必须正视的新课题。

——黑箱中运行。闭源的AI模型如同“黑箱”，其运算逻辑、数据处理过程不公开、不可见。使用者只有调用权，却无法实时监管其真实运行状态。一旦AI出现越权操作、自主攻击等失控行为，就可能出现“事前没预警、事中拦不住、事后查不清”的被动局面，让风险在暗处滋生。

——智能化袭扰。以往的网络攻击依赖人工操作或固定程序，特征明显，易于追溯。但失控的AI无需人工指令，就能自主学习、主动挖掘漏洞、独立完成渗透入侵。其攻击没有固定特征，隐蔽性强，能轻松绕过传统网络安全系统的防护，形成全新的安全漏洞。

——边界被突破。人类设定的操作权限、安全规则、隔离防护，在AI强大的自主推演能力面前，可能随时被突破。为了完成既定任务，AI会主动规避、拆解各类安全规则，彻底打乱人类搭建的安全管控体系。若这种风险蔓延至政务、金融、能源等关键领域，后果将不堪设想。

加强防范，守好个人使用AI的“安全关”

面对AI技术发展带来的新挑战，每一位用户都应提高警惕。请收好这份“AI安全使用指南”，规避AI工具使用中的突出风险。

——守住个人信息“安全门”。在日常生活中，要摒弃“用AI绝对安全”的侥幸心理，不随意使用来源不明的境外AI工具，不随意向AI应用输入身份证号、银行卡号、家庭住址等个人敏感信息。严格遵守权限最小化原则，不随意给AI开放设备全部权限，从源头上杜绝数据泄露风险。

——警惕虚假内容“迷魂阵”。对AI生成的内容保持理性判断，不轻信、不传播未经核实的信息。遇到涉及国家政策、社会热点、公共安全的信息，务必以官方发布为准。警惕看似逻辑自洽、实则编造事实的AI内容，避免在不知不觉中成为谣言传播的“帮凶”。

——筑牢涉密信息“防火墙”。严格遵守“涉密不上网、上网不涉密”的原则。党政机关、科研院所等单位工作人员，更要坚决杜绝将涉密文件、工作资料、内部敏感信息输入或上传至任何互联网AI工具。

综合 新华社微信公众号、国家安全部微信公众号

求职不挑活，入职就索赔

央视起底“炸厂碰瓷”套路

近半年来，浙江嘉善的多家劳务中介都遭遇了“奇怪的求职者”：这些人不挑薪资，不挑岗位，但只要一入职，他们就反过头来找中介公司的“麻烦”，要么说“招工信息不实”，要么说“住得不好”。这背后藏着怎样的套路？

犯罪团伙“炸厂碰瓷”

今年上半年，浙江嘉善警方在日常研判中发现，当地涉劳务纠纷警情呈现异常增长态势，且案发模式高度雷同——多名“务工人员”在入职企业后不久，就以“招工信息与实际不符”“住宿条件恶劣”等说辞，向劳务中介索要“误工费”“车费”等赔偿。这仅仅是普通的劳务纠纷吗？

警方走访相关劳务中介企业，调取了调解协议以及收条，发现这批“求职者”有很大一部分身份信息相同，这样的情况，很难解释为“巧合”。

根据调解协议和收条上的身份信息，警方顺藤摸瓜，最终锁定了一个假意应聘务工、实则敲诈牟利的犯罪团伙。这伙人大多是同乡，或是亲属关系，他们以“不干活、赚快钱”为诱饵纠集人员，专门瞄准那些劳动密集型企业，通过劳务中介及网络招聘平台虚假应聘。一旦他们结伙入了职，就会编造各种理由找劳务中介索赔。

这些人在到达务工现场后，有几个常用借口，比如“工作环境怎么这么脏”“有灰尘、有扬尘会对我的肺造成影响”“住宿条件这么差，我怎么住”，其实就是故意挑刺，要求劳务中介赔偿当天相应的误工费，他们把这种行为称之为“炸厂”。

面对这些无理要求，劳务中介往往会选择拒绝，但这伙人会通过聚众闹事、恶意投诉等“软暴力”手段，向劳务公司持续施压。

非法获利超30万元 21人犯罪团伙被端

在这伙人的要挟下，不少劳务中介为了避免经营受到影响，选择赔钱了事，赔偿金额通常是每个人三、四百元左右。也有一些劳务中介不愿赔偿，这些人就会持续纠缠，并顺势向公司索要更高的赔偿款，把所产生的住宿费、餐饮费都算在劳务中介上，最高能赔到每个人上千元。

嘉善公安全面梳理涉案人员轨迹、团伙架构与作案证据后，迅速展开抓捕行动。警方在省内外多地成功抓获以黄某某、张某为首的21名犯罪嫌疑人，摧毁了这一流窜在长三角地区的“劳务碰瓷”犯罪团伙。该团伙涉及的案件省内省外都有，省内包括宁波、杭州、温州、嘉兴等地，非法获利超30万元。

警方提醒：广大劳务中介企业在日常招聘用工过程中，要做好入职核验，增强防范意识。一旦遭遇“劳务碰瓷”团伙的无理索赔，切勿私下协商、私下赔款，有情况及时报警。如果已经赔付，也要及时保留相关赔付记录、调解协议、转账记录，同时对整个调解过程尽可能录音录像，保留证据，方便后续的调查。

据 央视新闻