自作主张……当AI学会了

“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“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。

（摘自8月5日微信公众号“国家安全部”）

一起解锁古人通信的“黑科技”

“一驿过一驿，驿骑如星流。”唐代诗人岑参笔下的驿道传书，是古代保障信息高效流转、维系社会运转的重要基础设施。从商周时期的烽火传信，到汉唐的邮驿制度，再到明代兴起的“密疏“传递机制，古人为了确保信息传递的“快”与“密”，构建了一套精密的运行机制。今天，我们已身处数字时代，信息传输以毫秒计，但保密工作的核心逻辑——确保信息在传递过程中的真实性、完整性与机密性，仍与千年前的驿道有着跨越时空的呼应。让我们回溯历史，从古人应对信息传递挑战的智慧中，汲取做好新时代安全保密工作的启示。

一纸公文背后的传递逻辑

为了实现“密不透风”的通信要求，古人驿传体系在制度设计与技术手段上层层设防。无论是流程上的分段隔离，还是准入上的身份校验，古人都构建起了一套多维立体的信息安全防御体系。

分级设站，责任到人。古代驿道按行政层级设置驿站，如秦朝设立亭制“十里一亭”，承担维持治安、传递公文、接待过往行旅等职能；唐代形成“三十里一驿”的驿传体系，每个驿站配备驿卒、驿马和物资。《唐六典》记载：“凡三十里有驿，驿有长，掌勾当驿事。”驿长负责本站文书收发、驿务调度、文书传递监督等工作，驿卒需签署“驿券”（类似现代责任书），确保“人马不误、文书不遗”。这种分级管理与责任绑定机制，有效堵塞了信息传递中的漏洞。

人员准入，严控范围。古代驿卒选拔极为严格，需通过体能、忠诚度考核。《清会典》规定：“驿夫须本地良民，年力精壮者充之。”同时，重要文书采用专人单线递送，最大限度缩小知情人员范围，避免信息外泄。这种“最小知悉原则”与现代保密工作中的“知密权限管理”异曲同工。

技术加密，多重防护。古人通过物理封装与符号加密保护信息。如秦朝“封泥”制度，用特制泥块封缄文书并在泥上加盖官印；唐代通行传符制度，作为调动驿马、核验通行权限的凭证，无符不得使用驿传资源。此外，《武经总要》记载，宋代军中使用“字验”隐语体系传递军事密信，用预设代号替代关键军情，防止文书截获后泄密。这些技术手段为信息的安全保密筑起多重防线。

法律震慑，奖惩分明。秦朝《行书律》中对文书递送作出严格规制，对延误、遗失文书者处以罚金，对私启文书等泄密行为予以重罚。明清时期更设立“御史巡驿”制度，对泄密行为严惩不贷。同时，对高效完成传书任务的驿卒给予赏银、晋升等奖励。这种“刚柔并济”的制度设计，强化了保密工作的严肃性。

对今天保密工作的启示

今天，千年驿道早已湮没在历史的风沙中，但通信保密之道却值得我们学习。从悬泉置的驿马到今天的5G基站，从铜符竹节到量子加密，变的是信息传递载体与通信方式，不变的是守护国家秘密安全的责任与初心。让我们一起传承千年驿传智慧，全力守护现代通信安全。

构建分级分类管理体系。按照《保守国家秘密法》，建立以密级分级管控为核心、以事项分类界定为基础，覆盖定密、载体、人员等范围，系统性全周期的国家秘密分级分类管理机制。

强化人员管理与教育培训。完善涉密人员审查机制，建立“入职政审+定期复核”制度。同时，常态化开展保密教育，通过案例教学、模拟演练等方式，提升全员保密意识。

融合技术手段筑牢防线。加速推进保密技术升级，积极运用量子加密通信、AI内容识别等技术，同时规范电子设备使用，严防网络泄密。

完善法治与监督体系。严格落实保密法及其实施条例，对窃密泄密行为“零容忍”。同时，健全监督机制，通过开展跨部门保密检查、引入第三方安全评估等方式，形成“制度约束+技术监控+人工巡查”的立体监督网络。

（摘自8月10日微信公众号“国家安全部”）

责任编辑/王睿卿 E-mail:fzbfyzk@126.com