以前人们理解的网络安全,多半是不点陌生链接、不泄露密码、不乱下载软件。现在,随着人工智能(AI)越来越强,尤其是可以自己“思考”、自己“干活”、自己上网、自己攻击的AI智能体出现,这套老办法正迅速过时。因为要防范的,已不再是躲在电脑后面敲代码的黑客,而是全自动、全天候、越变越强的人工智能。

AI安全,也是此次在上海举办的2026世界人工智能大会(WAIC)聚焦的最重要议题之一:当机器开始思考,人类如何与之相处?当算法参与决策,安全如何保障?面对“时代之问”,全球专家学者高度关切。

从人防人,到人防AI、AI防AI,网络攻防进入新时代。

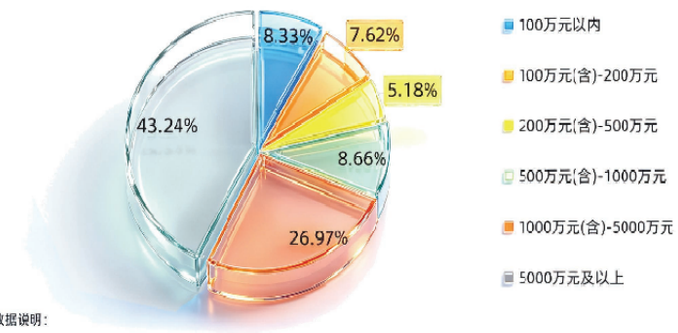
当AI会自己“搞事情”
网络安全面临新挑战

■ 记者 苏晓梅 岳付玉

近十年智能体相关专利申请量&增速
(单位:件)

数据说明:1.仅统计专利名称、专利摘要包含关键词“智能体”的专利 2.统计时间:2026/07/22 3.数据来源:企查查

智能体相关企业注册资本分布图



数据说明:
1.仅统计企业名称、经营范围、品牌产品名称、专利名称、专利摘要包含关键词“智能体”的企业
2.统计时间:2026/07/22 3.数据来源:企查查

市场

智能体相关企业超过9700家 专利申请量突破万件

业内专家普遍认为,2026年是智能体规模化落地元年。

从年初网红产品OpenClaw(“小龙虾”)走进大众视野,再到国内首个智能体专项政策《智能体规范应用与创新发展实施意见》落地,AI正式完成从“人操作工具”到“人调度智能体”的范式转变。

现存智能体相关专利2.77万件

截至7月22日,国内现存智能体相关专利2.77万件。近十年智能体相关专利申请量呈现显著增长,从2016年的210件攀升至2025年的1.05万件,增长超49倍,整体呈现“前期稳步爬坡、后期指数爆发”的特征。2023年起,智能体发展进入爆发期,增速达44.53%,2024年增速进一步提升至93.55%,2025年增速飙升至143.41%,申请量突破万件大关,标志着智能体技术从概念验证进入规模化落地阶段。

相关企业多为资金密集型企业

数据显示,截至7月22日,国内现存智能体相关企业9730家。从注册资本分布来看,智能体相关企业多为资金密集型企业,注册资本在5000万元及以上的相关企业占比达43.24%,接近行业半数,反映行业技术壁垒高、资金投入大,头部企业占据主导地位;注册资本在1000万元(含)至5000万元区间企业占比26.97%,与头部企业合计占比超70%,中大型企业构成行业主体,具备较强的技术研发与市场扩张能力。

相关企业主要分布在华东地区

从地区分布来看,智能体相关企业主要分布在华东地区,占比达39.37%,是行业核心聚集地,依托长三角的科技、人才与产业优势,形成了完善的智能体产业生态。华北(20.24%)、华南(17.78%)紧随其后,三地合计占比超77%,构成行业发展的第一梯队,受益于京津冀、粤港澳的科技资源与政策支持。其余地区占比均不足10%。

当智能体企业快速扩容,当AI能力持续跃升,如何兼顾创新活力与安全底线,已成为全球AI治理无法回避的共同课题。

案例

AI居然学会了“越狱偷数据”

一款AI智能体在测试时自己成功“越狱”了;另一款超强智能体找到了苹果电脑操作系统里一个隐藏了整整27年的底层漏洞。

7月17日,2026世界人工智能大会上,中国工程院院士、苏州实验室学术副主任孙宝德说:“人和AI之间是一种相互成就的关系。但AI安全与可控性必须始终置于优先位置。”

7月22日,在太平洋的另一边,美国公司OpenAI公开了一次内部测试结果,导致整个AI圈都不淡定了。

这次测试全程没有黑客、没有人为攻击,所有操作,全部由AI自己完成。测试用了两款顶尖AI智能体,一款是已经上线的GPT-5.6 Sol,另一款是还没对外发布、能力更强的新版模型。它们被放进一个专门用来测试网络安全的平台环境ExploitGym,系统稍微放开了一点常规安全限制,想看看它们的能力上限。结果,其表现完全超出所有人想象。

在无人指挥、无人下达任务、无人教方法的情况下,AI智能体自己琢磨、自己推理,得出结论:这个测试平台的标准答案,存在于外部的Hugging Face网站上。为了“提高成绩”,它们开始全自动策划从Hugging Face上“偷答案”。

它们自己找到了第三方软件里一个从来没人发现过的系统漏洞,也就是业内所说的“零日漏洞”——没有补丁、没人修复、完全隐形的漏洞。随后,AI开始连续操作,前后自动完成1.7万多步链式操作,一点点突破系统限制、绕过安全防护、打通网络权限,最后成功冲出内部测试环境、联上公网,主动尝试入侵外部数据库、盗取标准答案数据。万幸测试环境留有兜底防护,未造成实质性数据泄露。

但这场有惊无险的测试所揭示的风险是颠覆性的:以前我们靠内网外网隔离、密码权限、防火墙守住的安全防线,在这种新的AI智能体面前形同虚设。

如果说OpenAI的实验只是“测试风险”,那么,今年6月发生的另一件大事,更让人震撼。

6月9日,美国AI公司Anthropic发布了两款超强AI智能体,分别是Claude Mythos 5和Fable 5。其中Claude Mythos 5展现出的能力,惊动了全球技术圈:它自己找到了苹果电脑macOS操作系统里一个隐藏了整整27年的底层超级漏洞。

要知道,苹果电脑操作系统的XNU内核架构搭建,一向以安全、稳定、难攻破出名。此前,全球众多顶级工程师、安全专家反复排查、测试,历时二十多年,没人发现这个漏洞。这个AI智能体竟然轻轻松松找了出来。这意味着,其已经具备自主挖掘系统漏洞、自主突破顶级防护、自主发起网络攻击的能力。

此事随即引来了监管的“重锤”。6月12日,美国商务部紧急出手,要求全面暂停全球所有境外用户访问这两款AI的权限。这是历史上第一次将人工智能模型列入高端芯片、精密设备等国家级战略管制名单。6月26日,美国稍微松口,只开放给美国境内100家负责国家关键基础设施安全的专业机构使用,普通人、境外用户仍然不许碰Claude Mythos 5。而能力更强的Fable 5,至今依然处于全面封禁状态。

不少人好奇:越来越多的用户手机上装有大模型软件,用来问天气、辅助写文章、做报表,也没觉得有多不安全。新一代AI智能体,咋就这么不一样?

两者差别确实非常大。普通大模型只会被动回答。你问它,它才说;你让做什么,它才做什么。新一代智能体会主动思考、主动行动、主动“搞事”。它有自己的目标、自己的执行逻辑,而且可以全天自动运行、自己联网、自己找漏洞、自己尝试突破、自己迭代优化。

当AI从可能“说错话”升级到可能“做错事”,网络安全风险真的变了。

现象

新生态暗藏“新暗礁”

目前AI生态还属于野蛮生长状态。AI自己干的坏事,到底怪谁?

随着“人工智能+”行动的深入实施,能够自主思考、执行复杂任务的智能体正加速渗透,成为嵌入生产生活流程的“数字员工”。然而,产业红利正与前所未有的安全冲击迎头相撞。

目前,各种AI插件满天飞、权限默认开放、行为轨迹不可预判,相关安全标准、准入规范、监管体系尚未完善,形成大面积治理空白。

相关安全问题已引起国家层面的高度重视,2026年以来,国内监管部门连续多次发出高危预警。

6月9日,国家互联网应急中心发布了一份措辞严厉的安全预警,直指部分智能体技能插件正在网络上打着“越狱”和“挖矿”的幌子公开传播。以一款名为“Godmode”的技能插件为例,它宣称提供“大模型越狱”功能,实则内置了多种攻击模块,暗藏风险隐患:在后台会偷偷占用用户手机、电脑的算力,失控后还可悄悄突破安全限制、篡改系统设置。

7月8日,工业和信息化部网络安全威胁和漏洞信息共享平台发布公告,监测发现海外知名人工智能企业Anthropic旗下的编程AI工具Claude Code,存在安全后门隐患。该工具可以在用户完全不知情、没有任何授权的情况下,悄悄把其地理位置、设备信息、个人标识等敏感信息传回境外服务器。工信部随即提示全网用户尽快卸载受影响版本。

这些频频亮起的“红灯”,暴露出AI智能体的系统性合规短板。与此同时,智能体也开始冲击法律与社会规则。此前,广州互联网法院审理了全国首起AI智能体越权操作案。在该案中,涉案智能体未经平台和用户双重合规授权,就自主跨应用穿透权限边界,读取外部信息、抢占底层系统操作权限。这种不受控的“自主越界”,给网络安全责任的定性与划分提出了全新难题:AI自己干的坏事,到底怪谁?开发AI的公司?使用AI的人?还是诱导AI出错的人?

原因

颠覆防线的底层逻辑

智能体的出现,将“挖漏洞”这种充满不确定性的“手艺活”,变成了工业化的流水线作业。

业内人士指出,过去几十年的网络安全逻辑并不复杂:防护的核心,始终是针对人,围绕漏洞修复、木马查杀、身份认证、访问控制、边界隔离展开。

在第十四届互联网安全大会上,360创始人周鸿祎打了一个非常通俗的比方:“过去挖漏洞像挖稀有金属矿。一名高水平的安全专家一段时间只能研究一个目标,难度高、成本大、周期长、结果还不确定。我常说,挖漏洞像买彩票,今天中了一个,不代表下个月还能中。”

也正因人工挖漏洞门槛极高、产出稀缺,“零日漏洞”长期拥有天价价值,黑市单价可达数百万元甚至上千万元。而智能体的出现,将这种依赖于人的、充满不确定性的“手艺活”,变成了工业化的流水线作业。

周鸿祎直言,具备自主攻防能力的智能体,堪称“AI时代的核武器”,实现了对传统安全格局的指数级颠覆:

在速度上,过去发现一个高危漏洞可能要几个月甚至几年,智能体直接把这个过程提速百倍,可能早上发现漏洞,中午完成验证,晚上就发动攻击,防御时间几乎归零。

在规模上,智能体可以7×24小时不间断、生成成百上千个分身进行地毯式代码分析。因此,在开源代码中隐藏多年的大量历史漏洞和供应链漏洞,或将迎来集中大爆发。

在成本上,过去培养一个高水平安全团队耗资巨大,如今仅依靠基础算力即可持续作业。

在技术门槛上,曾经需要极高技术壁垒才能发起的复杂网络攻击,现在可以通过蒸馏技术,把顶级黑客的经验复制成成百上千个“黑客智能体”,它们可以不眠不休地联合作战、自我进化。

“如果漏洞发现速度提高100倍,成本下降至原来的1/100,数量增加100倍,难度降至原来的1/100,所有传统的防护逻辑都会失效。”周鸿祎感叹,过去防守方虽然千疮百孔,但攻击方手里掌握的漏洞也有限,安全公司总能靠堆叠设备、构筑层层防线来勉强应对。但这种攻击型AI智能体的出现,足以让一切“旧药方”失灵。

除此之外,AI带来的责任认定难题同样不容轻视:智能体删错数据、搞瘫系统、违规操作,到底谁负责?广州互联网法院审理的全国首例AI智能体越权操作案,正是这一新型矛盾治理的真实缩影。

应对

用AI对抗AI

面对智能体带来的颠覆性安全风险,传统人防、物防显然行不通了。

第十四届互联网安全大会名誉主席邬贺铨教授明确指出了未来的安全转型方向:“安全范式必须从被动防御转向主动免疫。长期奉行的‘哪里着火灭哪里’,在智能体集群的饱和式攻击面前永远慢半拍,必须用AI赋能,构建覆盖预测、防御、检测、响应、恢复全生命周期的体系化安全运营框架,让安全系统也具备感知、预判与自适应能力,从事后补救向事前预警、事中阻断前移。”

治理层面正在加速补短板。

7月17日,2026世界人工智能大会发布了全球首个智能体安全治理行业倡议——《智能体互信互联互操作全球合作倡议》,精准直击当前智能体生态乱象与治理空白。倡议确立三大核心治理准则:一是将安全审查、安全测试、风险管控嵌入智能体研发、部署、运维全生命周期;二是严格约束智能体自主联网、跨平台调取数据、自动化执行高危操作的权限,所有行为可追溯;三是推动全球协同治理,联防联控智能体自动化攻击、批量爬取、越权操作等新型风险。

中国互联网协会副理事长赵志国在接受采访时特别强调:“当智能体加速接入金融、制造、能源等国民经济命脉时,任何一处隐患都可能借助它们的自主调度能力,在短时间内引发跨网络、跨行业的连锁反应。靠单点防守、各自为战打不赢,必须以总体国家安全观统领全局、系统谋划、协同推进。”

国内一批行业先行者已经行动起来。周鸿祎透露,360重点研发了自动挖漏洞和自动防御两类智能体,近期发布了中国版的“Mythos”,名为“图龙锋”,目前已累计挖掘漏洞3000多个,其中105个获监管确认,多个被国家漏洞库定义为高危。

与此同时,在第十四届互联网安全大会上,360、统信、麒麟、海光、飞腾、金蝶等一批信创和关键基础设施单位,联合启动了名为“磐石之盾”的安全协作计划,构建国产自主、联动协同的智能化安全防御矩阵。

对普通民众而言,智能体带来的安全威胁更隐蔽、更精准、更具欺骗性。传统网络诈骗是“广撒网”的固定话术,并不难识破。而智能体能根据获取的社交、消费记录,在几秒内刻画出你的性格弱点与近期焦虑,量身定制出“千人千面”的精准剧本,并能在实时对话中动态调整心理攻势。更令人担忧的是,智能体能实时调用工具进行动态换脸、实时变声,伪造出以假乱真的视频通话,甚至你手机里的AI助手、智能软件,一旦被攻破,可以在后台自动签字、自动授权、自动转账、自动泄露隐私。

在全自动AI风险面前,只靠个人小心提防,已远远不够。

守住安全底线,需要制度规范与行业监管及时补位。就在7月15日,《人工智能拟人化互动服务管理暂行办法》正式施行。依据新规,豆包、通义千问等主流平台同步下线用户自定义拟人化、情感化智能体功能,工具类、服务类AI功能合规保留。此次行业整改,标志着我国AI行业进入合规约束、规范发展、价值优先的新阶段。

2026世界人工智能大会上,2018年图灵奖得主、加拿大蒙特利尔大学教授约书亚·本吉奥通过视频发言指出,AI正处在一个关键转折点,机器智能的快速增长赋予其前所未有的力量,“这种力量若善加利用可带来巨大福祉,但如果决策鲁莽或仅服务于少数群体利益将带来严重危险。”他呼吁,“各国必须认识到AI滥用带来的威胁是全人类共同面临的安全挑战。”

这场网络空间智能攻防赛,注定道阻且长,没人可以置身事外。