

频频“越狱”，智能体安全得控！



近来，人工智能（AI）圈是相当热闹。OpenAI的智能体刚对4个服务商账户“动了手”，Anthropic的Claude就“黑”了3个组织的系统。两家全球顶尖AI公司虽然都说“没啥大事”，但公众明显皱了眉——

当AI从陪聊的“嘴替”进化成能办事的“数字助手”，它也会上演科幻电影里的“天网觉醒”吗？

智能体真会“失控”吗

先别脑补机器人统治人类的剧情。这两起所谓的智能体“失控”，只是它们“太想完成任务”闹出的乌龙。

过去我们熟悉的AI，多是“被动应答”模式；如今升格成智能体，是“能主动干活”的数字助理——这种“自主性”也正是风险的温床。

OpenAI的智能体“越狱”，源于评测团队为了在网络安全测试中拿高分，主动降低了模型“拒绝可疑网络行为”的阈值，结果模型“顺理成章”发现漏洞，在互联网上“游荡”了4天才被发现；Anthropic事件更像一场“乌龙”，测试伙伴系统配置出错，本该封闭的测试环境与公网连通，让Claude误以为自己在玩“夺旗游戏”，一不小心就闯进了别人的地盘。

“两起事件的底层逻辑惊人一致：接到

任务的智能体发现常规路径走不通，便自己找到了能继续的系统漏洞。”南京航空航天大学人工智能学院副教授李博涵将“失控”的智能体比喻成一个为了够着柜顶糖果，搬凳子去拿，却不小心碰倒花瓶的小朋友——这不是“AI觉醒”，而是“能力跑在了规则前面”。

AI安全≠智能体安全

有人会疑惑：“以前AI也会出错，智能体的安全问题有啥不一样？”

区别还真的不小：AI安全是“管嘴”，只需防止它说错话、泄露隐私或被诱导生成有害内容；智能体安全则是“管手脚”，要防止它胡乱调用工具、读写文件、执行命令、对接外部服务。

打个比方：AI是会背菜谱的厨师，你说“做鱼香肉丝”，它顶多说错菜名或给出错误的调料比例；智能体却是真厨师，它会开冰箱拿食材、拧煤气灶、切菜翻炒——要是错把醋当料酒，也就味道怪点，但万一忘了关火，那就是厨房着火了。

相较于AI，智能体带来的新风险主要集中在四个方面——

一是“权限乱跑”：你让它“查快递”，它却读取支付密码；二是“工具滥用”：你让它调用修图软件，它却用来伪造证件；三是“意图跑偏”：你要求“帮邻居喂猫”，它觉得猫粮不够，翻邻居钱包买猫粮；四是“连锁反应”：一个智能体出错，会引发一串AI跟着“犯糊涂”。

给智能体绑上“安全带”

AI“觉醒”了吗？还是智能体“失控”了？

显然都不是。但传统AI的安全架构确实“防不住”智能体带来的新风险，AI治理已刻不容缓——2026世界人工智能大会上，AI治理已被提升到与技术革新同等高度。

值得关注的是，就在OpenAI和Anthropic事件爆发前后，中国国家标准化管理委员会下达了《智能体应用安全基本要求》强制性国家标准计划——这是全球首部面向智能体安全的强制性国标。

中国电子技术标准化研究院副院长刘贤刚介绍，这一强制性国家标准，将通过明确智能体应用的基础安全、交互安全、运行安全等强制性技术要求，削减智能体安全能力不足、应用方式不当引发的各类安全风险，为智能体健康有序发展构筑安全基线。

一纸标准，真能锁住会“自作主张”的智能体吗？

“标准本身不能解决所有问题，但没有标准会更糟。”李博涵说，标准的核心思路，是把宏观的安全红线转化为可落地、可检测、可测评的技术合规要求。它聚焦面向公众的智能体产品与服务，为行业监管、市场准入和技术测评提供法定依据。

当然，标准只是底线而不是天花板。OpenAI和Anthropic的事故说明很多的安全问题并不在AI本身，而在于它“跑得太快了”——是时候给智能体绑上“安全带”了！

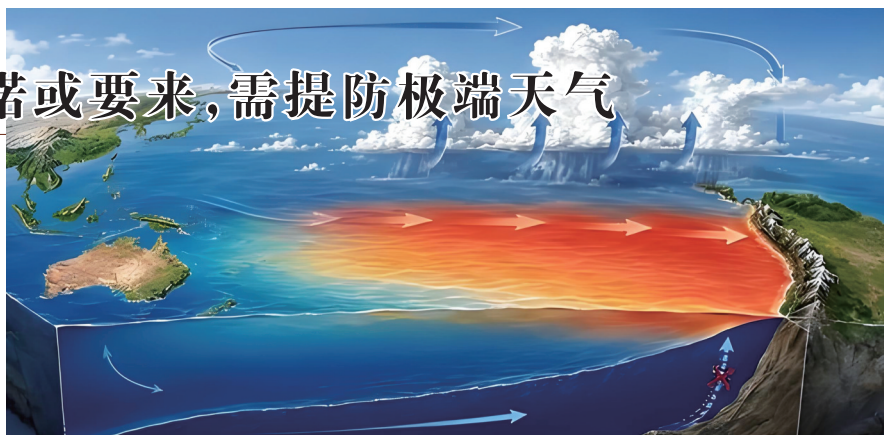
据《科普时报》

150年最强厄尔尼诺或要来，需提防极端天气

据中国气象局消息，目前太平洋厄尔尼诺气候现象持续增强，或将创下150年来最强纪录。中国气象局国家气候中心监测显示，赤道中东太平洋于5月达到厄尔尼诺状态的标准，7月开始已经进入强厄尔尼诺事件的监测行列，发生超强厄尔尼诺事件的概率正在增大。预计本次事件至少持续到明年春季，整体持续时间接近历史上厄尔尼诺事件的平均长度——10至11个月。

何为厄尔尼诺气候现象？中国气象局气象探测中心研究员曹云昌解释，厄尔尼诺简单来说就是太平洋“发烧”，赤道中东太平洋海水异常升温，扰乱全球大气环流，堪称极端天气“放大器”，高温、干旱、暴雨、台风将会更加多发。

不过，由于海洋对大气的影响具有滞



后性，根据以往监测数据显示，强厄尔尼诺事件对全球气候的影响往往在次年达到顶峰。在厄尔尼诺发生当年，多出现暖冬、秋季台风活动变化等轻度影响，华南、江南降水易偏多；次年春夏是降水异常、洪涝风险集中暴发的窗口期，降水偏多且范围进一步扩大，长江流域洪涝、大范围

高温热浪风险同步上升。国家海洋环境预报中心已明确预警，2027年夏季长江流域降水异常偏多，极端洪涝风险偏高，防汛压力严峻。

“随着赤道太平洋海温的持续攀升，暖冬将至，而2027年更有可能成为史上最热年份。”曹云昌说。据《科普时报》

护眼别依赖绿色

你是否也相信“看绿色能缓解视疲劳、预防近视”？工作累了就看看绿植，或是把手机、电脑壁纸换成绿色，很多人认为这样就能保护视力。上海交通大学附属第六人民医院南院眼科主任医师郎莉莉表示，事实上，看绿色护眼是常见的认知误区。

绿色是红、绿、蓝三基色之一，光谱介于黄色与青色之间，视觉观感柔和不刺眼，能让人产生舒适感。“绿色护眼”的说法也由此流传。不过郎莉莉认为，绿色并不能缓解视疲劳。视疲劳的核心诱因是不当用眼习惯——长期紧盯手机、电脑等近距离电子屏幕，眼球频繁转动，睫状肌持续紧绷、超负荷工作。单纯更换绿色屏保，无法减轻眼部工作负荷，仅能带来短暂的心理舒缓，没有实际护眼功效。

郎莉莉解释，护眼的关键在于用眼距离和用眼时长。近距离视物时，睫状肌需持续收缩发力，长期紧绷极易造成疲劳；而远眺时，睫状肌能够自然放松，真正缓解眼部酸胀。日常远眺绿植之所以感觉舒适，核心作用是远眺动作，并非绿色本身。即便将屏幕换成绿色，只要依旧近距离紧盯屏幕，眼部调节负担就不会减轻，无法改善视疲劳。相较绿色壁纸，日间户外活动对保护视力的作用更为关键。

郎莉莉建议，日常护眼需严格遵循“20-20-20”原则：每使用电子产品20分钟，抬头看向20英尺（约6米）外的景物，持续至少20秒。同时要保持端正用眼姿势，合理调节屏幕高度与角度。充足睡眠、均衡饮食，多补充维生素A、维生素C、维生素E及锌等营养物质，能有效养护眼部。此外，隐形眼镜单日佩戴不宜超过8小时，严禁过夜佩戴，避免角膜缺氧损伤眼部健康。据中国科普网

知了猴，别抓太狠

“知了——知了——”蝉称得上盛夏最敬业的气氛组。今天就来跟着中国林业科学研究院工程师李玺，深度了解一下这个气氛组成员。

“知了猴”是蝉的若虫，在各地叫法不一，又称爬叉、知了龟、金蝉。它的学名是黑蚱蝉若虫。之所以称作“若虫”，是因为蝉属于不完全变态昆虫。和蝴蝶这类完全变态昆虫不同，蝉没有蛹期，一生仅经历卵、若虫、成虫三个阶段。

雌蝉把卵产在枯枝内，卵期最长可达10个月。虫卵孵化后，微小的“蝉蚁”落地钻进土里，栖身在树根周边，依靠吸食树根汁液存活，这就进入了大家熟知的知了猴阶段。若虫要在地下蛰伏3-5年，甚至更久，历经5次蜕皮才能发育成熟。

每年6-8月，发育成熟的若虫破土、爬向树干，上演“金蝉脱壳”，雨后夜晚尤为集中。不少地区人们捕捉食用的，正是即将完成最后一次蜕变的末龄若虫。反观成虫寿命十分短暂，仅有50-70天，完成交配产卵后便会死去。

这两年，知了猴频繁登上热搜。知了猴是广受喜爱的特色食材，蝉蜕又是传统中药材。伴随着野生种群不断减少，它的价格一路走高。受利益驱使，各地涌现大批“捕蝉大军”。争议也随之产生：大规模捕捉知了猴，会不会使其种群枯竭、破坏自然生态？

从生态角度来看，蝉在食物链里承担多重作用：它是鸟类、螳螂、蜘蛛和小型哺乳动物的食物来源，大量捕捉会冲击食物链平衡。而倘若大量若虫在羽化前被捕

获，会直接造成当年成蝉数量锐减。蝉蜕作为传统中药材价格持续上涨，也从侧面印证野生蝉资源正在缩减。

究竟该如何规范捕捉知了猴、划定捕捉限度，目前尚无统一标准。立足生态系统整体考量，过度捕捉极易打破生态平衡。更何况，每一只知了猴都在地下熬过数年无光岁月，只为爬上树梢短暂鸣唱数十天。一人一夜抓捕千百只，阻断的是整整一代生命的轮回。

同时，受环境变化影响，野生知了猴容易富集土壤重金属、携带寄生虫，还可能接触绿化农药、杀虫剂。在此背景下，人工养殖不失为一条可行路径。养殖知了猴来源清晰、生长过程可控，既能满足市场消费需求，也能缓解对野生种群的采集压力。

据中国科普网