

频频“越狱”，智能体还安全吗



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。”8月4日,南京航空航天大学人工智能学院副教授李博涵接受科普时报记者采访时,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底线。

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》报道

乳状云是特大暴雨的预兆吗



●天空中出现的乳状云。

近日,“乳状云 特大暴雨”话题冲上热搜。出现乳状云,是否预示着特大暴雨即将来临?

“乳状云”并不是一种独立的云,而是一种特殊的云附属特征,根据《国际云图集》,可以附着在多种云体上,如积雨云、高积云、卷积云等云层上。中国气象局气象探测中心研究员曹云昌解释,乳状云确实和强对流天气有关,雷雨、暴雨、冰雹、龙卷风出现时,都有可能伴随它现身。“但是,乳状云本身并不直接产生降水,因此,它不是前兆,而是暴雨尾声。”

仔细观察会发现,我们平常看到的云都是向上长的,但乳状云看起来却是一朵朵、一袋袋的云团倒

新华社记者 刘续 摄

挂在云层下方,或白色或灰色,像倒扣的棉花糖。“这与它的形成原理有关。”曹云昌介绍,云层底部的冷空气下沉,周围的暖空气上升,冷暖气流剧烈碰撞、相互拉扯,在浮力和重力的作用下,就形成了这种独特的乳状云结构。简单说,乳状云是冷暖气流在云层底部“打架”留下的痕迹。

其实,乳状云出现时,强对流天气大多已结束,就算还在下雨,雨势也正在快速减弱。“大家别再被误导了!看到乳状云,只能说明附近有强对流天气,并不意味着你头顶即将倾泻特大暴雨。”曹云昌说。

据《科普时报》报道

暴马丁香缘何被称为“植物冰箱”



●左图为暴马丁香,右图为暴马丁香开花。作者供图

提起丁香,大多数人想到的是春天开花的低矮灌木。但丁香属有约30种,并非全都长这样。有一种春末夏初开花的丁香,叫作暴马丁香,人们称它为“北方木本花卉之王”。

暴马丁香又名暴马子,原产中国,是木犀科丁香属里典型的乔木,株高可达4米至15米。不同于常见的低矮灌木丁香,它身姿挺拔、长势高大,是丁香家族名副其实的“大高个儿”。

作为晚花品种,暴马丁香的花期从5月末延续至7月初。盛花时,枝头缀满舒展蓬松的白色圆锥花序,素雅洁净,花心暗藏黄色花药,丰沛花蜜吸引蜂蝶传粉。它的香气浓烈,远超一般灌木丁香;雄蕊也极有辨识度,细长的花丝伸出花冠,远比花瓣要长。而它的近亲“北京丁香”的雄蕊藏于花冠之内,不凑近看很难察觉。凭借这一点,就能轻松将它们区分开。

暴马丁香是丁香家族中最耐严寒、分布最北的树种,广泛生长在我国东北、华北及西北高原。高原气候严苛,菩提树难以存活,而暴马丁香不仅耐寒耐贫瘠,还长着酷似菩提叶的心形叶片,因此被尊称为“西海菩提树”。

除了颜值出众、适应性强,暴

马丁香的木材还被称为天然“植物冰箱”。它无需制冷,木材中含有的丁香酚、抗氧化物质与挥发油,能抑制细菌繁殖、延缓食物氧化,自带清雅木香,同时,暴马丁香材质耐腐蚀、不变形,可形成洁净稳定的储存环境。在东北,人们常用这种木材制作碗橱、茶叶筒与衣柜,既能保鲜饭菜、锁住茶香,又能防虫防蛀,无需樟脑丸便可保护衣物。民间更有传说,唐代运送杨贵妃爱吃的荔枝,也曾靠它来保鲜。

暴马丁香全身是宝,实用价值高。它的树皮、树干及枝条均可入药,具有清肺祛痰、止咳平喘、消炎利尿功能。知名止咳药“琴(qín)暴红”中的“暴”,指的就是暴马子皮。暴马丁香的花苞和嫩叶可制茶,茶汤清亮、清香微甜,有清凉解暑、镇咳祛痰的作用。同时,它还是“暴马子桑黄”的天然寄主,其上生长出的桑黄,具有较高的药用价值。

据《科普时报》报道

AI“社交零食”未必能解情感“饥饿”

随着生成式AI技术的爆发,一种“能说会聊”的AI悄然走红,成为很多人的“社交零食”,数百万用户与它们建立了深厚情谊。但如果你正打算将它视为情感的避风港,希冀靠它填补感情“饥饿”,那你可能要失望了。

美国斯坦福大学科学家研究表明,向聊天机器人袒露心声、掏心掏肺,不仅不会获得幸福,反而可能会让一些人更感孤寂。他们考察了用户与Character.AI平台上AI之间的互动,收集了1131人的调查数据,其中244人提供了完整的聊天记录。参与者标明了使用聊天机器人的

首要动机,例如,提高工作效率、寻求情感支持、让生活更便利等,并用自己的话描述与AI形成的关系。随后,他们借助GPT-4o、LLaMA 3-70B和TopicGPT等多款AI工具,深入分析了用户如何与聊天机器人互动,以及这些互动和用户主观心理状态之间的关联。

科学家重点关注三个方面:为什么使用聊天机器人,互动有多频繁深入、愿意倾诉多少敏感信息,包括从情绪困扰到药物使用,甚至是否有自杀念头。同时评估用户的线下社交网络,比如能无话不说、秉烛夜谈的亲友的数量。

有意思的是,用户嘴上说的和心里想的并不一样。明确把AI当作主要动机的不到12%,但超过50%的用户使用“朋友”“伴侣”这类词来描述与聊天机器人之间的关系,而八成以上提供的聊天记录都试图从聊天机器人那里寻求情感和社会支持。

那些互动频繁、对使用聊天机器人感到有意义的人,幸福感得分的确更高。但在现实社交圈本就狭小的人群中,频繁使用聊天机器人反而幸福感更低。此外,更愿意向AI坦白私密心事的人,幸福感也可能更低。这与人际交往

中“倾诉增进亲密”的效果正好相反。

这可能是因为聊天机器人既不会像真人那样分享自己的心事,也难以真正识别并恰当回应情绪化的对话。这就好比用AI递给内心“饥饿”的人一包“社交零食”。这类零食虽然暂时可以“解饿”,填补人们心中的孤独和焦虑,但缺少长期情绪健康所需的真正营养。而且,若不警惕,这些“社交零食”甚至可能酿成恶性循环:社交圈越小的人越容易被AI吸引,越沉迷就越少参与现实社交,孤独感也愈演愈烈。

据《科技日报》报道

山珍 杀手 良药 野生菌的“善变”面孔

近期,广西、四川、云南等地群众因采摘误食野生菌引发的中毒事件时有发生,为安全敲响了警钟。

云南气温回升、雨季来得早,相比其他地区,野生菌长势好、种群也多,素有“野生菌王国”的美誉。统计数据显示,目前,世界已知野生食用菌2500余种,云南就有900余种;而在我国已知有毒的500余种野生菌中,云南占有200余种,其中20余种含剧毒。

“见手青中毒在临床最常见。它会给患者带来幻觉。但这种体验未必美妙,不少患者看到的是恐怖景象。”云南省第一人民医院急诊内科医生宫冰介绍。见手青是一类伤氧化后出现靛蓝色反应的牛肝菌的统称,常见的有兰茂牛肝菌、华丽新牛肝菌等。它们属于条件可食,只要烹饪得当,毒素会被高温分解,但不建议自行采食。

值得警惕的是,在众多有毒野生菌中,“头上戴帽(菌盖)、腰间系裙(菌环)、脚上穿靴(菌托)”的致命鹅膏,则是误食中毒事件中的“头号杀手”,可造成严重肝损伤。

“千足虫”:别怕,我不咬人

入夏之后,雨水增多、空气潮湿,走在公园步道上,常能见到成群“千足虫”出没,长短粗细各不相同,身子一节一节的,既形似“条形码”,又酷似蜈蚣,看着让人心里发怵,不少市民担心其可能伤人。

北京林业大学林学院副院长石娟介绍,北京常见的“千足虫”,大多是北京小直形马陆,偏爱树下阴凉潮湿、腐殖质丰富的环境。它既不是昆虫,也不是蜈蚣,不会主动咬人、蜇人,大家无需过度恐慌。

以腐殖质为食的马陆,虽然长得“不讨喜”,却是自然界关键的生物分解者,助力生态系统物质循环。近

期,马陆数量增多,跟北京这段时间多雨、潮湿的天气有关。若是天气干热少雨,它们会躲进落叶堆、泥土缝隙,数量明显变少。

常见的马陆主要有两种:黄缘条马陆,也就是北京小直形马陆,体长3-4厘米,躯体扁黑,两侧镶着金黄色边,属于本土物种;燕山蛩(qióng),又称金环马陆,体形偏大,身长可达10-12厘米,躯体呈圆筒黑褐色,大部分体节后缘带有金黄色横纹。



●7月1日,云南省楚雄彝族自治州人民医院蘑菇中毒临床研究室,药剂师李吉品在整理标本柜。

有人分不清马陆和蜈蚣,其实,一句口诀就可以快速区分:爬得慢、身体圆、受惊蜷缩是马陆;跑得快、身体扁、每节一对足、前端有毒爪是蜈蚣。

碰到马陆如何正确处理?石娟建议:户外遇见马陆,可用树枝、纸片或扫帚轻轻移开,切勿徒手抓、用脚踩,这是因为马陆受惊、受压时,会从身体腺体释放防御性液体,少数人接触后可能出现灼热、泛红、瘙痒,甚至起水泡。如果皮肤不慎接触到马陆分泌物,用肥皂水冲洗,不要反复揉搓;若马陆的分泌物溅入眼睛,立即用清水或生理盐水冲洗10-15分钟。一旦皮肤持续红肿刺痛、起水泡或视物不适,应及时就医。

身体扁、每节一对足、前端有毒爪是蜈蚣。

碰到马陆如何正确处理?石娟建议:户外遇见马陆,可用树枝、纸片或扫帚轻轻移开,切勿徒手抓、用脚踩,这是因为马陆受惊、受压时,会从身体腺体释放防御性液体,少数人接触后可能出现灼热、泛红、瘙痒,甚至起水泡。如果皮肤不慎接触到马陆分泌物,用肥皂水冲洗,不要反复揉搓;若马陆的分泌物溅入眼睛,立即用清水或生理盐水冲洗10-15分钟。一旦皮肤持续红肿刺痛、起水泡或视物不适,应及时就医。

据《科普时报》报道

当汽车底盘有了“大脑”

高速公路上,一辆新能源汽车正以120公里的时速巡航。突然,左前轮发生爆胎。驾驶员还没来得及做出任何反应,方向盘已完成自行微调,车身在零点几秒内轻微晃动后迅速恢复稳定。车速平稳下降,最终安然停靠应急车道。

“以前,这是科幻电影的画面。现在,智能底盘技术正逐步将它变为现实。”西南交通大学汽车研究院副院长杨明亮告诉科普时报记者。传统底盘的转向与制动系统依赖物理连接,方向盘通过金属杆直接拉动车轮。智能底盘的创新在于用电子信号取代机械传动,电线不仅比铁杆更轻,更关键的意义在于,底盘由此成为可编程的硬件。“方向盘与车轮之间的对应关系不再被物理锁定,工程师可通过软件设定不同的驾驶风格:低速泊车时灵敏轻巧,高速行驶时沉稳精准,切换只在参数之间。”杨明亮介绍说。

但仅靠线控还不够。杨明亮说,传统底盘是被动的,当车轮压到坑洼时,车载减振器才会去压缩吸收。“智能底盘与车辆的摄像头、雷达等感知系统深度融合,路面尚未被车轮触及,底盘已提前‘看见’。从‘踩到再反应’到‘看见就准备’,底盘的反应逻辑发生了质变。”

更关键的是,智能底盘引入车辆运动控制系统(VMC),将驱动、制动、转向、悬架等所有执行单元纳入统一协调。车辆运动控制系统则像一个“中央大脑”,能够在毫秒级时间内感知异常并做出全局决策。高速爆胎时,车辆传感器在毫秒内感知,系统同步指挥制动与转向协同动作,在驾驶员介入前就完成了一系列稳定车身的操作。

此外,智能底盘对道路保护也有积极意义。车辆超重,对路面的压力就越大,造成破坏程度就越重。智能底盘的线控悬架尤其是线控主动悬架,可吸收部分冲击能量,减小路面承受的压力值。

任何车辆使用时间长了,底盘的性能都会发生衰减。传统底盘的控制程序在出厂时已经固化,无法感知和补偿这些变化。智能底盘通过传感器持续采集行驶数据,借助大数据与人工智能算法,识别性能衰减趋势,并以云端推送的方式更新控制软件。“不管汽车用多长时间,底盘都能保持操控和舒适性能。”杨明亮说。

电动化是上半场,智能化是下半场。未来消费者选购电动汽车时,除了关注续航里程和芯片算力,或许还需要多问一个参数:这台车的底盘,运行的是哪个版本?

据《科普时报》报道