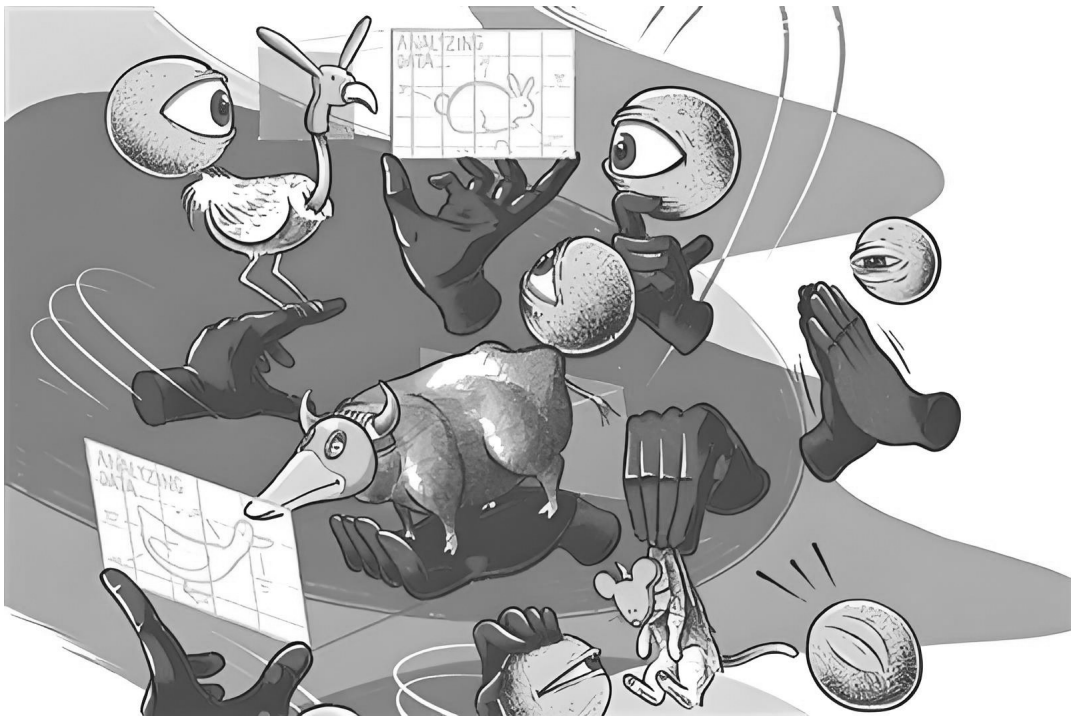
数据“中毒”会让AI“自己学坏”



AI系统在学习过程中,如果输入了错误或误导性数据,可能会逐渐形成错误认知,做出偏离预期的判断。 图片来源:英国《新科学家》网站

在一个繁忙的火车站,监控摄像头正全方位追踪站台的情况,乘客流量、轨道占用、卫生状况……所有信息实时传输给中央人工智能(AI)系统。这个系统的任务是帮助调度列车,让它们安全准点进站。然而,一旦有人恶意干扰,比如用一束红色激光模拟列车尾灯,那么摄像头可能会误以为轨道上已有列车。久而久之,AI学会了把这种假象当作真实信号,并不断发出“轨道占用”的错误提示。最终,不仅列车调度被打乱,甚至还可能酿成安全事故。

澳大利亚《对话》杂志日前报道称,这是数据“中毒”的一个非常典型的例子。AI系统在学习过程中,如果输入了错误或误导性数据,可能会逐渐形成错误认知,作出偏离预期的判断。与传统的黑客入侵不同,数据“中毒”不会直接破坏系统,而是让AI“自己学坏”。随着AI在交通、医疗、媒体等领域的普及,这一问题正引起越来越多的关注。

AI“中毒”的现实风险

在火车站的例子中,假设一个技术娴熟的攻击者既想扰乱公共交通,又想收集情报,他连续30天用红色激光欺骗摄像头。如果未被发现,这类攻击会逐渐腐蚀系统,为后门植入、数据窃取甚至间谍行为埋下隐患。虽然物理基础设施中的数据投毒较为罕见,但线上系统,尤其是依赖社交媒体和网页内容训练的大语言模型中,它已是重大隐患。

据英国《新科学家》杂志报道,2024年,互联网出现了一个标志性事件,即AI爬虫的流量首次超过人类用户,其中OpenAI的ChatGPT-User占据了全球6%的网页访问量,它本质上是ChatGPT的“上网代理”,在用户需要实时信息时替他们访问网站。而Anthropic的ClaudeBot更是长期大规模抓取网页内容,占到13%的流量。

互联网上的大量内容正被AI模型不断采集、吸收,用于持续训练。一旦有人故意投放有毒数据,比如篡改的版权材料、伪造的新闻信息,这些大规模采集的爬虫就可能把它们带进模型,造成版权侵权、虚假信息扩散,甚至在关键领域引发安全风险。

版权之争中的“投毒”反击

随着AI爬虫的大规模抓取,许多创作者担心作品被未经许可使用。为了保护版权,创作者采取了法律和技术手段。

面对旷日持久的版权拉锯战,一些创作者转向技术“自卫”。美国芝加哥大学团队研发了两款工具。名为Glaze的工具可在艺术作品中加入微小的像素级干扰,让AI模型误以为一幅水彩画是油画。另一款工具Nightshade更为激进,它能在看似正常的猫的图片中植入隐蔽特征,从而让模型学到“猫=狗”这样的错误对应。通过这种方式,艺术家们让自己的作品在训练数据中成为“毒药”,保护了原创风格不被复制。

这种反击方式一度在创作者群体中风靡。Nightshade发布不到一年,下载量便超过一千万次。与此同时,基础设施公司Cloudflare也推出了“AI迷宫”,通过制造海量无意义的虚假网页,将AI爬虫困在假数据的循环中,消耗其算力和时间。可以说,数据投毒在某些领域已经从一种反击手段,演变为版权与利益之争中的防御武器。

去中心化成为AI的防护盾

这种局面让人警觉。创作者的数据“投毒”是为了保护原创,但一旦同样的技术被用于大规模制造虚假信息,其后果可能比版权争议严重得多。

面对这种隐蔽的威胁,研究者正在探索新的防御手段。在美国佛罗里达国际大学的Solid实验室,研究人员正着力用去中心化技术来防御数据投毒攻击。其中一种方法叫联邦学习。与传统的集中式训练不同,联邦学习允许模型在分布式设备或机构本地学习,只汇总参数而非原始数据。这种方式降低了单点中毒的风险,因为某一个设备的“坏数据”不会立刻污染整个模型。

然而,如果在数据汇总环节遭遇攻击,损害依然可能发生。为此,另一种工具——区块链正被引入AI防御体系。区块链的时间戳和不可篡改特性,使得模型更新过程可被追溯。一旦发现异常数据,可追根溯源,定位投毒源头。同时,多个区块链网络还能互相“通报”,当一个系统识别出可疑模式时,可立刻警示其他系统。

任何依赖现实世界数据的AI系统都可能被操纵。利用联邦学习和区块链等防御工具,研究人员和开发者正在打造更具韧性、可追溯的AI系统,在遭遇欺骗时能发出警报,提醒系统管理员及时介入,降低潜在风险。

来源:科技日报

昆虫尺度的软体机器人问世 “机器小强”强在哪儿

它从108米高的标志塔顶急速坠落,重重砸向地面,可就在片刻之后,这个仅有2厘米长、2克重的小家伙竟重新起身,若无其事地蹦跶起来,像极了“打不死的小强”……这不是科幻电影,而是西湖大学工学院姜汉卿实验室研发的新一代软体机器人。该研究团队提出了一种全新的电磁弹性体驱动机制,首次让昆虫尺度的软体机器人在复杂户外环境中实现完全自主运动,为未来小型化、无线化、高性能的智能机器人系统提供了全新解决方案。相关成果日前发表于国际学术期刊《自然·通讯》。

姜汉卿教授告诉记者:“在自然界,昆虫靠肌肉高效收缩爆发出惊人力量,但人类复刻这一奇迹却困难重重。传统机器人依赖笨重的电机和复杂零件,根本无法塞进昆虫般小巧的身体。而曾被寄予厚望的人工肌肉,又往往需要高压电或强磁场驱动,难以在户外自由施展。”

如何突破这一困境?

姜汉卿团队从昆虫肌肉的伸缩机制中获得灵感,创造出全新的电磁弹性体驱动机制。这个系统巧妙结合了磁力与弹性:利用弹性力和静磁吸力的平衡,来实现机器人类似肌肉收缩的运动。他们还设计了一个精巧的驱动系统,将其塞进了软体机器人不到2厘米长的小身板里。如今,只需不到4伏的低电压,线圈磁场便能让机器人像肌肉般高效收缩,爆发出高达210牛/千克的力量和60%的惊人形变,性能远超现有技术。

姜汉卿透露,这精密的“人工肌肉”暗藏玄机,其弹性体能如拉满的弓般储存能量,形成独特的“双稳态”甚至“三稳态”。这意味着,机器人完成动作后无需持续耗电就能保持状态。“举个例子,新一代软体机器人能耗低,只有56毫瓦,跟小LED灯差不多。它背着一枚8毫米长、4毫米厚、容量只有20毫安的小型板载电池时,可以持续工作一个小时。”

正是这项突破性技术,让一群“机器小强”被赋予多种运动模式,使其得以真正走出实验室,在复杂户外环境中大显身手。

——当它从高空坠向平地,抗摔打的本领立刻凸显价值。1.6厘米长的蠕动机器人从108米高空自由落体后,竟能毫发无伤地继续在尘土中匍匐前进。未来,在地震等重大自然灾害发生后,它可以被无人机投放落地,快速进入废墟深处,寻找被困人员位置并发出信号,成为生命搜救的“先锋”。

——当2厘米长的游泳机器人跃入水中,瞬间化身灵巧的“机器鱼”,在自然水体中自主巡游超1小时。未来,可以派小巧灵活的它检测水下环境或监测污染。

——全球最小的自主跳跃软体机器人在草丛间腾跃,指尖大小的身躯在崎岖地形中连续起跳,展现出令人惊叹的运动潜力。未来,它有望在复杂地形上感知环境、自主移动、躲进缝隙等。

未来,这些“机器小强”将挑战两栖运动与3D跨障能力,成为人类执行一系列科学任务、探索极端环境的得力助手。“或许在不久的将来,当紧急情况发生时,第一批抵达现场的‘使者’,就是这些蹦蹦跳跳、从天而降的‘机器小强’。”

来源:光明网