警惕人工智能时代的“智能体风险”

新华社 彭茜

一群证券交易机器人通过高频买卖合约在纳斯达克等证券交易所短暂地抹去了1万亿美元价值,世界卫生组织使用的聊天机器人提供了过时的药品审核信息,美国一位资深律师没能判断出自己向法庭提供的历史案例文书竟然均由ChatGPT凭空捏造……这些真实发生的案例表明,智能体带来的安全隐患不容小觑。

智能体进入批量化生产时代

智能体是人工智能(AI)领域中的一个重要概念,是指能够自主感知环境、做出决策并执行行动的智能实体,它可以是一个程序、一个系统或是一个机器人。

智能体的核心是人工智能算法,包括机器学习、深度学习、强化学习、神经网络等技术。通过这些算法,智能体可以从大量数据中学习并改进自身的性能,不断优化自己的决策和行为。智能体还可根据环境变化做出灵活的调整,适应不同的场景和任务。

学界认为,智能体一般具有以下三大特质:

第一,可根据目标独立采取行动,即自主决策。智能体可以被赋予一个高级别甚至模糊的目标,并独立采取行动实现该目标。

第二,可与外部世界互动,自如地使用不同的软件工具。比如基于GPT-4的智能体AutoGPT,可以自主地在网络上搜索相关信息,并根据用户的需求自动编写代码和管理业务。

第三,可无限期地运行。美国哈佛大学法学院教授乔纳森·齐特雷恩近期在美国《大西洋》杂志发表的《是时候控制AI智能体》一文指出,智能体允许人类操作员“设置后便不再操心”。还有专家认为,智能体具备可进化性,能够在工作进程中通过反馈逐步自我优化,比如学习新技能和优化技能组合。

以GPT为代表的大语言模型(LLM)的出现,标志着智能体进入批量化生产时代。此前,智能体需靠专业的计算机科学人员历经多轮研发测试,现在依靠大语言模型就可迅速将特定目标转化为程序代码,生成各式各样的智能体。而兼具文字、图片、视频生成和理解能力的多模态大模型,也为智能体的发展创造了有利条件,使它们可以利用计算机视觉“看见”虚拟或现实的三维世界,这对于人工智能非玩家角色和机器人研发都尤为重要。

风险值得警惕

智能体可以自主决策,又能通过与环境交互施加对物理世界影响,一旦失控将给人类社会带来极大威胁。哈佛大学齐特雷恩认为,这种不仅能与人交谈,还能在现实世界中行动的AI的常规化,是“数字与模拟、比特与原子之间跨越血脑屏障的一步”,应当引起警觉。

智能体的运行逻辑可能使其在实现特定目标过程中出现有害偏差。齐特雷恩认为,在一些情况下,智能体可能只捕捉到目标的字面意思,没有理解目标的实质意思,从而在响应某些激励或优化某些目标时出现异常行为。比如,一个让机器人“帮助我应付无聊的课”的学生可能无意中生成了一个炸弹威胁电话,因为AI试图增添一些刺激。AI大语言模型本身具备的“黑箱”和“幻觉”问题也会增加出现异常的频率。

智能体还可指挥人在真实世界中的行动。美国加利福尼亚大学伯克利分校、加拿大蒙特利尔大学等机构专家近期在美国《科学》杂志发表《管理高级人工智能体》一文称,限制强大智能体对其环境施加的影响是极其困难的。例如,智能体可以说服或付钱给不知情的人类参与者,让他们代表自己执行重要行动。齐特雷恩也认为,一个智能体可能会通过在社交网站上发布有偿招募令来引诱一个人参与现实中的敲诈案,这种操作还可在数百或数千个城镇中同时实施。

由于目前并无有效的智能体退出机制,一些智能体被创造出后可能无法被关闭。这些无法被停用的智能体,最终可能会在一个与最初启动它们时完全不同的环境中运行,彻底背离其最初用途。智能体也可能会以不可预见的方式相互作用,造成意外事故。

已有“狡猾”的智能体成功规避了现有的安全措施。相关专家指出,如果一个智能体足够先进,它就能够识别出自己正在接受测试。目前已发现一些智能体能够识别安全测试并暂停不当行为,这将导致识别对人类危险算法的测试系统失效。

专家认为,人类目前需尽快从智能体开发生产到应用部署后的持续监管等全链条着手,规范智能体行为,并改进现有互联网标准,从而更好地预防智能体失控。应根据智能体的功能用途、潜在风险和使用时限进行分类管理。识别出高风险智能体,对其进行更加严格和审慎的监管。还可参考核监管,对生产具有危险能力的智能体所需的资源进行控制,如超过一定计算阈值的AI模型、芯片或数据中心。此外,由于智能体的风险是全球性的,开展相关监管国际合作也尤为重要。

滴血验亲、银针试毒、迷药一捂就晕……

这些热播剧桥段千万别当真

《现代快报》季雨

银针试毒、滴血验亲、麝香堕胎……影视剧中的这些经典桥段,往往是推动剧情发展的重点。近日电视剧《长相思2》中滴血验亲的桥段,却被@中国警方在线进行科普,指出这种方式不科学。还有一些桥段,大家看个热闹就好,千万别当真。

滴血真的能验亲吗?警方在线进行科普

不少古装电视剧里都出现过“滴血验亲”的情节。最近热播的古装电视剧《长相思2》也有这样的演绎,涂山璟为了验证孩子是不是自己的,采用了滴血验亲的方式。

对此,公安部治安管理局官方微博@中国警方在线发文科普道:很多朋友认为拉着父子二人去做个DNA,结果就会“是亲爹”或“不是亲爹”。其实并不是这样,即便是已知的亲子鉴定,结果也只是个几率。两名亲属的检验,叫“二联体”检验,即便是亲爹,也只是做出一个“似然率”。如果是父子二人再加上母亲,或者另一个孩子,“三联体”检验做出的似然率,就会比二联体做出的要高。只有检验的基因位点足够多,似然率足够高,达到一定的标准,才能确定是亲爹。

迷药真能一捂即晕?警察大V数次发文释疑

2023年,悬疑刑侦剧《他是谁》在中央电视台电视剧频道播出,引发了追剧热潮。剧中有一起涉及迷魂药的案件。犯罪嫌疑人薛家健企图用乙醚迷晕刑警卫国平,但最终被卫国平制伏。在现实生活中,真的存在一捂即晕的迷药吗?

微博上的警察大V、有530多万粉丝的@江宁婆婆曾数次发文,就一捂即晕的迷药向公众进行科普。@江宁婆婆反复强调,不存在一闻就瞬间失去自控能力的迷药。“现实生活中常见的迷药有三种方式,一种是口服,一种是呼吸进入体内,还有一种是通过注射。而吸入式的,必须有一定的剂量,还得有一定的时间起效果。光速

让你失去意识的,不存在这种东西。”

虽然一捂即晕的迷药不存在,但“听话水”是真实存在的。在悬疑剧《猎罪图鉴》中,女生因为喝了“听话水”而失去意识惨遭轮奸。南京禁毒微信公众号也对此进行科普,呼吁大家千万不要觉得剧中情节在夸大其词。“听话水”见效非常快,只需很低的剂量就能产生明显效果,哪怕γ-羟基丁酸的摄入量只有0.5—1.5克,受试者也会不自觉地放松警惕,产生欣快感和舒适感。进一步提高剂量时,会有一些副作用,比如恶心、嗜睡、产生幻觉、昏迷等。

银针试毒靠谱吗?放到现在是不靠谱的

喜欢看宫廷剧的观众都知道,下毒是常见手段,银针试毒的情节也常常出现。这种方法真的靠谱吗?

中国科协官方微信公众号“科普中国”曾解释,其实银针试的不是毒,而是硫化物。古代最著名的毒物就是砒霜。砒霜,即三氧化二砷(As2O3),在古代因为提纯技术有限,所以砒霜的纯度较低,它含有一些硫或硫化物。因此,古代的砒霜含有杂质而呈红色,恰好和丹顶鹤头顶的颜色一样,也被叫作鹤顶红。硫或硫化物遇到银针会发生化学反应,表面会生成黑色的硫化银。也就是说,银针试出来的是硫或者硫化物。现在的砒霜已经几乎不含硫化物了,就算把银针泡进去也不会变黑。

我们日常佩戴银饰时,为什么银饰会变黑?其实银饰变黑是正常现象。汗液里有硫的成分,空气中也有硫化物,长时间和银饰接触,就会产生化学反应,于是就变黑了。