警惕人工智能时代的“智能体风险”

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。 新华社记者 彭茜

新研究：创作过度依赖AI可能导致雷同

新华社华盛顿7月15日电 美国《科学进展》杂志日前发表的一项新研究说，生成式人工智能（AI）的兴起可能使影视、文学、音乐等的创作变得容易起来，但如果创意产业过度依赖AI“编故事”，未来可能出现作品千篇一律的雷同感。

生成式AI可将简单的文本提示转化为相对复杂的音乐、文本、图像和视频等，但这类工具的广泛使用会对人类创作产生何种影响仍然未知。为了解生成式AI对短篇小说创作的影响，英国伦敦大学学院等机构的研究人员招募了近300名志愿者作为“作家”，展开一项在线研究。

这些志愿者并不是以创作为生的职业作家。研究人员评估了他们的先天创作力，然后将他们随机分为3组。所有志愿者被要求根据随机分配的公海探险、丛林探险和外星探险3个选题之一，创作一个8句话的小故事。

3组志愿者接受生成式AI辅助创作的程度不同。第一组无任何AI辅助；其他两组可选择利用AI获得一个3句话的初始创意；在这两个允许借助AI创作的组别当中，有一个组的志愿者可选择最多获得5个由AI产生的创意。创作完成后，志愿者们被要求以新颖性、情感特征等标准对自己创作的故事自我评估。此外，还有600名外部评审人员以相同标准来评估这些故事。

研究显示，接受生成式AI辅助有助于创作更有创意、更有趣的故事，这在一开始被测定为先天缺乏创作力的志愿者中尤其明显。例如，对于先天缺乏创作力的志愿者，获得5个AI提供的创意可使他们创作故事的新颖性提高10.7%，可使他们故事的趣味性提高22%。但从总体来看，与无任何AI辅助组相比，AI辅助组创作的故事看起来更相似，因为他们在创作时太过依赖AI提供的故事创意。

研究人员说，这就相当于创造了一种“社会困境”：生成式AI使人们更容易踏入写作领域，“降低门槛是好事”。但如果整体的艺术创新程度降低，那将是有害的。这项研究表明，人们必须开始思考，在工作中如何利用AI来获取最大益处，同时又保有自己的思考。

特朗普被提名为共和党总统候选人 万斯为副总统候选人

新华社美国密尔沃基7月15日电（记者 熊茂伶 刘亚南）美国前总统特朗普15日在共和党全国代表大会上获得足够多的党代表票，被正式提名为2024年美国总统选举共和党总统候选人。特朗普当天还宣布，已选择俄亥俄州联邦参议员詹姆斯·万斯作为他的竞选搭档。

美国国会众议院共和党籍议长迈克·约翰逊当天在大会上正式宣布，提名特朗普和万斯为共和党总统和副总统候选人。

在继2016年击败希拉里·克林顿、2020年败给现任总统拜登之后，这将是现年78岁的特朗普第三次代表共和党参加美国总统竞选。

当天早些时候，特朗普在其创建的社交媒体平台“真实社交”上宣布，他已选择万斯作为他的竞选搭档。

万斯生于1984年，2022年当选俄亥俄州联邦参议员，并于2023年1月宣誓就职。他曾是特朗普的激烈批评者，但此后成为了这位前总统的盟友。

特朗普13日在宾夕法尼亚州巴特勒市举行的竞选集会上遭“未遂刺杀”，右耳受伤。但他按照原计划于14日抵达威斯康辛州的密尔沃基，参加15日至18日举行的共和党全国代表大会。特朗普预计于18日正式接受提名。

7月15日，在美国密尔沃基，特朗普（前左）和万斯出席共和党全国代表大会。 新华社 发

梅措拉再次当选欧洲议会议长

新华社法国斯特拉斯堡7月16日电（记者 陈斌杰）欧洲议会16日在法国斯特拉斯堡召开全会，马耳他籍的萝伯塔·梅措拉经投票再次当选欧洲议会议长。

俄罗斯春季征兵15万人应征入伍

新华社莫斯科7月15日电（记者 江宥林）据俄罗斯国防部15日消息，今年俄春季征兵工作已经结束，共有15万人应征入伍并被派往俄武装部队和其他军事编队。

俄国防部当天在社交媒体上发布消息说，俄罗斯各征兵委员会于4月1日启动工作，应征入伍者派遣工作于4月15日开始。为保障军事运输，俄武装力量飞机15个航班、14个军事梯队、172个民航航班、多列客运列车以及军队公路运输车辆参与本次征兵工作。

按惯例，俄罗斯每年春季、秋季各征兵一次。2023年春季征召约14.7万人入伍，2023年秋季征召13万人入伍。

阿曼一清真寺附近发生枪击 至少4人死亡

新华社科威特城7月16日电（记者 尹炯）马斯喀特消息：阿曼警方16日在社交媒体上通报说，该国一清真寺附近发生枪击事件，目前已造成4人死亡、数人受伤。

阿曼皇家警察16日凌晨在一份简短声明中表示，枪击发生在阿曼东北部马斯喀特省凯比尔干河地区的苏丹卡布斯大清真寺附近，目前已造成4人死亡、数人受伤。

警方表示，已逮捕枪手并正在调查其动机。警方已采取必要的安全措施，并向死伤者及家属表示哀悼和慰问。

苏丹卡布斯大清真寺是阿曼最主要的大清真寺，也是一个著名的地标，可同时容纳2万人祈祷。

新研究尝试用疫苗改善心衰实验鼠心脏功能

新华社东京7月16日电（记者 钱铮）日本研究人员近日在美国学术期刊《循环》上报告说，他们发现心脏血管内皮细胞分泌的一种蛋白质可能诱发心力衰竭，针对这种蛋白质开发的疫苗能够改善心衰实验鼠的心脏功能，预防心衰发生。

除心脏移植外，心力衰竭目前尚无根治方法。日本东京大学等机构的研究人员通过动物实验发现，实验鼠心脏中老化的血管内皮细胞会分泌蛋白质IGFBP7，而这种蛋白质会抑制心肌细胞的线粒体代谢，导致心脏功能障碍，诱发心衰。他们还研发出了靶向蛋白质IGFBP7的疫苗，心衰实验鼠接种疫苗后，心脏功能得到了改善。

研究人员表示，这种疫苗疗法为治疗心衰提供了新的可能性。疫苗生产成本低，接种容易，副作用小，如能通过疫苗治疗心衰，将使更多患者得到救治。另外，给心衰发病风险高的人群接种疫苗，还可能预防心衰发生。

阿富汗北部车祸致17人死亡

新华社喀布尔7月16日电（记者 邹学冕 赵家淞）阿富汗官员16日说，一辆大客车当天早上在阿北部巴格兰省坠入山谷，造成17人死亡、34人受伤。

巴格兰省警方发言人希尔·艾哈迈德·布尔哈尼接受新华社记者电话采访时说，因司机驾驶不当，客车坠入山谷，死者中包括3名儿童，一些伤员情况危急。

阿富汗多年来局势动荡，道路设施建设缓慢，交通事故频发。

禽流感疫情持续在美蔓延 美疾控中心派专家组协助应对

新华社北京7月16日电 美国农业部、美国疾病控制和预防中心近日表示，禽流感疫情持续在美蔓延，出现奶牛和人感染禽流感病毒的病例。尽管如此，美疾控中心表示，目前禽流感对公众健康构成的风险仍较低。

美国农业部15日确认，俄克拉何马州奶牛样本中发现禽流感病毒，该州成为美国第13个在奶牛中发现禽流感病毒的州。

俄克拉何马州农业、食品和林业部门发言人李·本森说，今年4月，俄克拉何马州一家奶牛场怀疑其奶牛可能感染禽流感病毒，遂采集样本并于近期送到美农业部检测。

美农业部在其网站公布，俄克拉何马州有两个牛群报告禽流感病毒阳性。本森表示，这两个确诊阳性的牛群属于同一个奶牛场，目前奶牛已经完全康复。俄克拉何马州尚未收到其他牛群感染禽流感病毒的报告。

俄克拉何马州官方表示，该州已经为奶牛场工人提供了防护装备，并要求奶牛场针对禽流感病毒采取安全防护措施，但目前尚无针对奶牛的强制性检测措施。美国今年3月首次在奶牛中检测到禽流感病毒。此后，美国已在150多个奶牛群中检测到禽流感病毒。

科罗拉多州卫生部门14日表示，该州4名家禽养殖场工人确诊感染H5N1型高致病性禽流感病毒，另有1例疑似病例，样本已送到美疾控中心检测确认。这些感染者症状轻微，包括结膜炎和发烧、咳嗽、喉咙痛等，均未住院治疗。据介绍，这些工人在科罗拉多州东北部一家暴发禽流感疫情的蛋鸡养殖场负责屠宰家禽。

美疾控中心14日发表声明称，该机构已派遣一支由流行病学家、临床医生、兽医和工业卫生学家等组成的专家小组，前往科罗拉多州协助应对禽流感疫情。美疾控中心表示，目前科罗拉多州其他地区尚未出现禽流感病毒活跃度异常上升的情况。同时，针对病毒的基因组测序也正在进行中，该机构将关注禽流感病毒可能出现的突变，这将影响疫情风险评估结果。

近几个月来，H5N1型高致病性禽流感病毒一直在全球野生鸟类中传播，同时已出现感染家禽和哺乳动物的情况。美疾控中心表示，虽然禽流感对公众健康构成的风险仍较低，但人群如果接触受感染或潜在受感染的动物，那么受感染风险将更大，因此建议采取相应预防措施。

美国今年已报告多例与奶牛感染禽流感病毒相关的人感染病例。英国《自然》杂志网站今年5月曾刊发文章说，一些研究人员表示，美国在应对奶牛中出现的禽流感疫情时存在数据收集和报告不足的情况，这不利于评估禽流感暴发规模，也会阻碍防止病毒进一步传播的努力。 新华社 发

巴基斯坦一军营遇袭致8名军人死亡

新华社伊斯兰堡7月16日电（记者 蒋超）巴基斯坦军方16日说，该国西北部开伯尔－普什图省一军营15日遭恐怖袭击，8名军人死亡，发动袭击的10名恐怖分子全部被打死。

巴基斯坦三军新闻局发表声明说，10名恐怖分子15日凌晨企图冲入开伯尔－普什图省本努地区一处军营，他们驾驶一辆装满炸药的汽车撞向军营围墙引发爆炸，部分围墙倒塌、毗邻设施受损，并导致8名军人死亡。巴安全部队在交火中打死全部恐怖分子。

声明说，实施这起恐怖袭击的是名为“哈菲兹·古尔·巴哈杜尔集团”的组织，该组织在阿富汗活动。

巴基斯坦总理夏巴兹16日发表声明谴责此次袭击，并称赞巴安全部队及时采取行动挫败恐袭。

开伯尔－普什图省与阿富汗接壤，巴基斯坦塔利班等恐怖组织常年在此活动。该省也是巴军方重点开展反恐行动的区域。

新发现的月球洞穴或可为宇航员提供天然庇护

新华社北京7月16日电 一个国际团队日前在英国《自然·天文学》杂志发表研究成果说，他们在月球静海区域发现了一处熔岩管洞穴，这处洞穴以及其他类似的洞穴或可为宇航员提供天然庇护。

由意大利特伦托大学学者领衔的团队分析了美国航天局探测器获取的月球观测数据，他们发现在月球静海区域表面下存在一处熔岩管洞穴，所处位置距离“阿波罗11号”飞船的落月点约400公里。

雷达数据仅显示了洞穴的一部分。研究人员通过数据分析为这一熔岩管洞穴的一部分建立了模型。据估计，整个洞穴至少宽40米，长几十米，并且有入口。

熔岩管洞穴是一种特殊的洞穴类型，是熔岩在流动并固结过程中形成的中空管道。科学家此前推测月球上存在不少这类洞穴。

这项发现不仅是月球科研的新进展，也为人类探月带来新的可能性。月球表面环境非常严酷——部分地区表面温度可高达127摄氏度，另一些地区则低至零下173摄氏度，还有极强的宇宙射线和太阳辐射等。如果要长期深入探索月球，非常需要为宇航员建立安全的庇护所。

据研究人员介绍，未来这类洞穴有潜力发挥天然庇护所的作用，可能只需进行洞穴墙体加固或者搭建一些设施，就能帮助宇航员抵御宇宙射线和太阳辐射等并持续开展探索活动，这比在月球上修建全新的庇护基地显然简单很多。

阿富汗东部两省洪灾造成至少40人死亡

7月15日，在阿富汗东部楠格哈尔省苏尔赫罗德，居民查看洪灾损失。

阿富汗官员15日说，阿东部楠格哈尔省与库纳尔省当天因暴雨引发洪水灾害，已造成至少40人死亡、230人受伤。 新华社 发

本报地址：芜湖市红花山路20号 邮编：241003 区号：0553 编辑部：3831155 采访部：3117073 专副刊部：3836809 校对部：3839927 广告经营许可证：皖芜工商广字001号 广告热线：3813185、3815764 报刊公司：3852053 芜湖日报报业集团印务有限公司承印（地址：芜湖市红花山路20号） 每份零售2元