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。

7月6日，在上海举行的世界人工智能大会上，人们观看机器人跳舞。 新华社记者 王翔 摄

2023年11月2日，首届人工智能安全峰会在英国举行，一名参会者经过宣传展板。 新华社记者 李颖 摄

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。 （新华社北京7月16日电 记者彭茜）

科普

为什么花季的年龄却得了卵巢早衰？

小玲27岁，结婚3年，婚后一直在备孕，但事与愿违。她很惆怅、焦虑，于是到生殖中心门诊就诊。医生追问病史发现，小玲一年前开始出现月经不规律，周期20-90天，睡眠质量较差，晚上睡觉易出汗，情绪不稳定。经检查，她被诊断为卵巢早衰。她非常困惑，为什么花季的最佳生育年龄却会得卵巢早衰呢？今天我就给大家科普一下。

● 什么是卵巢早衰？

卵巢早发性功能不全（POI）：指女性在40岁以前出现卵巢功能减退，主要表现为月经异常（闭经、月经稀发或频发）、卵泡刺激素（FSH）水平升高大于25 U/L、雌激素水平波动性下降。卵巢早衰是卵巢早发性功能不全的终末阶段。简而言之，就是本该30-40岁的卵巢却活成了50岁的样子。由于卵巢衰退的时间提前，所以对身体的影响很大，造成的危害也比自然绝经更严重，尤其是骨质疏松和心脑血管问题。更残酷的是：卵巢功能衰竭是不可逆转的！未来只能用持续服用激素治疗到平均绝经年龄，来替代卵巢功能。

● 卵巢早衰有哪些表现？

（1）月经改变：可先后出现月经频发或稀发、经量减少、闭经。（2）雌激素水平下降的表现：患者可有潮热出汗、生殖道干涩灼热感、性欲减退、骨质疏松、情绪和认知功能改变、心血管症状等。（3）不孕、不育：在卵巢储备减退的初期，由于偶发排卵，仍有5%左右的自然妊娠机会，但流产和胎儿染色体异常的风险明显增加。（4）其它：有的表现为乳房萎缩、阴毛和或腋毛脱落、外阴阴道萎缩等。

● 卵巢早衰有哪些病因？

（1）遗传因素：主要包括染色体异常和基因突变。（2）免疫因素：常见的自身免疫性疾病有桥本甲状腺炎、系统性红斑狼疮、类风湿关节炎等都可引起卵巢功能损伤。（3）医源性因素：因女性患有疾病所采取的化疗、放疗及手术治疗均可导致医源性POI的发生。（4）环境因素及不良生活方式：环境中存在的某些化学、物理及生物因素会对女性的卵巢功能产生有害影响而损伤卵巢功能。经常吸烟喝酒，长时间久坐不动，经常熬夜，过度减肥等不良生活习惯及压力过大等，都是导致卵巢早衰的原因。

卵巢不可能永葆青春，大多数女性的绝经年龄是在48-50岁，那要怎么保护卵巢防止它过早衰老呢？

正确的卵巢保养方法就在生活的点点滴滴当中，一是保持好心情，好心态尤为重要；二是合理均衡饮食。不长期吃素，少吃甜食，少吃冰冷辛辣刺激性食物，少吃动物脂肪；三是科学运动。运动能调节内分泌，保持激素分泌平衡，促进血液循环，加快新陈代谢，从而有效延缓卵巢衰老；四是保持充足睡眠。研究发现，人体所需激素在夜间分泌最为旺盛，女性经常熬夜容易导致体内激素环境发生改变。所以，女性朋友要养成良好的睡眠习惯，千万不要熬夜。

对于女性来说，适龄婚育是最好不过的，对于卵巢也是最自然的保养。

（海南医科大学第一附属医院生殖科副主任医师 朱娟）