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。

2023年11月2日，在英国布莱奇利园，一名参会者经过首届人工智能安全峰会的宣传展板。新华社 发

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 （新华社北京7月16日电）

数字告诉你 中国游泳运动员接受兴奋剂检查『千锤百炼』

东京奥运会期间，中国游泳奥运冠军张雨霏曾经在受到外媒质疑时表示：“中国运动员接受兴奋剂检查的次数是全世界最多的！”此言不假，仅从世界泳联公布的2023年兴奋剂检查数据就能看到，“蛙王”覃海洋、张雨霏和李冰洁三位“泳士”接受的兴奋剂检查最频繁，一人的检查次数相当于四到五位外国奥运冠军的总和。

有“蝶后”之称的东京奥运会两金得主张雨霏去年总共接受了43次国内国际兴奋剂检查，包括27次赛外和16次赛内检查。与她同项目的加拿大女子100米蝶泳奥运冠军麦克尼尔的检查总数则是11次，包括7次赛外和4次赛内检查。瑞典奥运冠军舍斯特伦的数字是10次，含7次赛外加3次赛内。另一位美国奥运奖牌获得者胡斯克的检查次数为10次，含6次赛外加4次赛内。21岁的沃尔什在6月举行的美国奥运选拔赛中爆冷打破女子100米蝶泳世界纪录，还是福冈游泳世锦赛女子4×100米混合泳接力冠军成员，而她去年竟然没有接受过一次兴奋剂检查！

覃海洋在福冈世锦赛上将男子蛙泳三个项目的世界冠军收入囊中，刷新200米蛙泳世界纪录，并荣膺世界泳联年度最佳男子游泳运动员，2023年他的兴奋剂检查数字达到46次，包括24次赛外和22次赛内。而英国“蛙王”皮蒂的全年检查数字为9次，全部来自赛外飞行检查。

擅长女子中长距离自由泳的“棒棒冰”李冰洁则接受了42次兴奋剂检查，含24次赛外和18次赛内。同项目的美国“多金王”莱德茨基的检查数字为11次，包括6次赛外和5次赛内。

女子100米和200米自由泳高手杨浚瑄去年接受了27次兴奋剂检查，女子200米和400米自由泳世界纪录保持者、澳大利亚名将蒂特姆斯只有7次。

在今年2月多哈游泳世锦赛上改写男子100米自由泳世界纪录的潘展乐去年共有29次兴奋剂检查，包括17次赛外、12次赛内。美国奥运冠军德雷塞尔去年只接受了7次检查，全部是赛外；前世界纪录保持者、罗马尼亚的波波维奇去年共接受了10次检查；澳大利亚“飞鱼”麦克沃伊的次数为7次。

此外，2023年男子200米个人混合泳奥运冠军汪顺总共接受了22次检查，400米混合泳世界纪录保持者、法国人马尔尚是12次；男子100米仰泳世锦赛冠军徐嘉余“领走”26次检查，美国奥运冠军墨菲为12次；新科女子100米蛙泳世界冠军唐钱婷进行了21次检查，美国名将莉莉·金的数字为10次。

很多其他外国名将们的兴奋剂检查数字也“平淡无奇”：创造今年女子200米蝶泳最好成绩并改写400米混合泳世界纪录的加拿大17岁新星麦金托什去年进行了16次检查。今年打破女子100米仰泳世界纪录的美国选手里根·史密斯去年接受了10次检查，她的200米蝶泳目前世界排名第二。原100米仰泳世界纪录保持者、澳大利亚的麦基翁则有11次。

据了解，这份来自世界泳联的兴奋剂检查“成绩单”是各国（地区）运动员接受国内国际赛内和赛外检查的数据汇总，巴黎奥运会前夕更是加大了对中国运动员的检查力度。面对频繁的交叉式兴奋剂检查，中国游泳运动员高度配合，而且团结一心、斗志昂扬。正如覃海洋所言：“让实力打破一切质疑！”

巴黎奥运会游泳比赛将于7月27日开赛，共进行35个泳池项目的比拼。中国游泳队派出31人的阵容，目前在法国多维尔进行赛前适应训练。

此前，就23名中国游泳运动员2021年食品污染致阳性事件，世界反兴奋剂机构进行了调查，瑞士独立检察官埃里克·科迪尔在提交的报告中表示，世界反兴奋剂机构对此事的处理“无偏颇”，当时没有对中国反兴奋剂机构的决定提起上诉“毫无疑问是合理的”。世界泳联15日表示，其特别设立的反兴奋剂审查委员会也做出结论性报告，表明世界泳联在此事的审查中“没有违规、处置不当或隐瞒”。

新华社记者 周 欣 马向菲 李 嘉
（新华社北京7月16日电）