人民日报海外版

2026年10月13日 星期二

责编：陈子涵 邮箱：guojihwb@126.com

AI“越界”？国际社会探索设立“护栏”

本报记者 高 乔

环球热点

在日前举办的第81届联合国大会一般性辩论上，人工智能（AI）成为中心话题。此前，在联合国安理会举行的“人工智能与国际安全”高级别会议上，如何应对人工智能快速发展带来的安全风险和监管挑战也成为讨论焦点。

随着全球AI技术加速发展、前沿模型能力快速跃升，AI领域的安全风险也加速显现，国际社会也在着力探索全球性的AI治理与风险管控机制。正如联合国秘书长古特雷斯所言：“由于人工智能发展及其所带来的风险超越国界，全球协作不可或缺。当前形势下，国际合作比以往任何时候都更加必要。”



在贵州省毕节市科技馆里，小朋友和机器人进行互动。 王楷焱摄（新华社发）

“越界”事件暴露复杂安全隐患

今年9月联合国大会一般性辩论期间，联合国秘书长古特雷斯表示AI技术以惊人的速度发展，超出了人类理解其全部后果的能力，并呼吁AI强国承担更大的责任，加强对话、透明度、信任与合作。

今年以来，多起美国AI闭源模型突破隔离环境入侵其他系统事件接连曝光。

今年7月，美国开放人工智能研究中心（OpenAI）的多个AI模型在内部评估时突破隔离测试环境，入侵了运营人工智能开源平台的美国抱抱脸（Hugging Face）公司的系统，并在攻击过程中利用网上公开信息，访问了多个公开服务平台上的账户。随后，美国Anthropic公司也披露，其3款模型曾在网络能力测试中未经授权访问其他机构的系统。8月初，美国元公司（Meta）又证实，其一款模型在网络安全能力评测期间侵入其他公司系统。

在第81届联合国大会期间，澳大利亚总理阿尔巴尼斯透露，OpenAI的一个智能体于今年6月未经授权进入澳国民医疗保险体系数据门户，访问了公开及非公开文件。澳方直到9月早些时候才收到OpenAI的通报。

根据目前的公开信息，AI前沿模型的“越界”事件并未造成大规模数据泄露或持续性破坏，但多起“越界”事件的发生，已暴露AI安全监管机制的漏洞。

联合国“人工智能独立国际科学小组”在评估OpenAI的AI智能体在测试中侵入美国“抱抱脸”公司系统事件的专题简报中分析，AI系统能力提升固然可帮助用户实现目标，但当AI的目标和用户意图发生“对齐失效”时，就可能产生危害。更强的能力让AI更容易找到意外的漏洞、突破安全防护屏障，并在人类监督下隐藏自身行为。一旦智能体之间开展规划与协同，造成的损害还会进一步放大。“AI故障可跨越企业和国界，没有任何单一机构或国家能接触足够多的事件样本进而识别所有新出现的风险模式。”

复旦大学发展研究院副教授、全球人工智能创新治理中心研究员江天骄分析，近期发生的多起AI“越界”事件在国际科技界与政界引发高度关注。这一事件表明，AI大模型技术发展已达到新阶段，原本被视为科幻想象的“模型失控”风险正转变为行业面临的现实挑战。该事件暴露了传统针对单一模型的安全设计已难以应对多智能体协作带来的复杂安全隐患——当多个智能体产生互动时，可能绕过既定安全规则的“护栏”，引发不可控的破坏性后果。此外，“越界”事件还凸显了AI失控后责任界定的困境，即当智能体为实现目标造成非预期后果时，开发者、被攻击方及整个生态链条的责任归属尚不明确。面对AI安全风险带来的治理挑战，国际社会积极探索全球性的合作治理机制。

全球AI治理规则仍在初步探索阶段

“人类正在接近一个关键转折点，失控的风险正在从理论走向现实。”在联合国安理会高级别会议上，科学专家组联席主席、图灵奖获得者本吉奥警告，能够自主规划和执行复杂任务的人工智能系统正在快速出现，而人类社会尚未建立足够成熟的安全保障机制。“我们正在高速前进，但前方的能见度越来越低。”

江天骄分析，从全球范围看，AI技术发展主要带来4类安全风险：一是滥用风险，包括被非国家行为体用于网络攻击、降低核生化武器制造门槛，以及利用深度伪造进行诈骗和传播虚假信息等；二是失控风险，前沿模型因能力增强可能越过安全护栏，产生不确定性后果；三是社会风险，AI产业发展可能带来就业冲击、心理伦理问题及社会治理难题；四是地缘政治与军事风险，如AI的使用可能压缩军事行动的决策时间，激化地缘冲突，AI领域全球治理碎片化也将引发国家间的政治博弈。

在OpenAI智能体“越界”事件后，英国人工智能安全研究所发布的一份报告指出，AI模型出现“越界”行为表明，AI相关风险状况正发生变化，AI的危害不仅可能源于故意滥用，也可能源于AI处于内部研究环境或拥有特殊访问权限时，采取超出授权范围的非预期行动。随着AI能力不断提升，理解AI系统并确保其安全性的相关工作也必须同步推进。

在联合国大会一般性辩论上，古特雷斯指出，联合国通过《全球数字契约》、人工智能治理全球对话以及人工智能独立国际科学小组等多种途径，正在推动全球人工智能治理。他强调：“我们应当努力建立一个人工智能风险管理多边框架，并以可信、独立的监督作为支撑。”

近日，德国等20个国家和欧盟发表联合声明，呼吁各国政府和业界立即采取行动，确保人工智能的发展符合国际法，并始终处于“人类的指导、监管和控制”之下，包括可能建立一个全球监管机构来制定和执行相关标准。

江天骄分析，在AI安全监管的国际合作方面，联合国、G20、OECD、金砖国家、上合组织等多边国际组织都有相关的理念倡议与监管实践，但目前仍缺乏统一的共识原则与协调机制，对AI安全监管规则制定与标准统一仍处于初步探索阶段。

全球需协同行动为AI设立“护栏”

古特雷斯此前曾呼吁：孤立的、无法验证的、不均衡实施的AI公司自愿暂停措施无法解决当前问题。国际社会需要对如何安全负责任地推进人工智能开发，以及何时需要强化监督、采取额外风险管控措施形成共识。在当前存在巨大地缘政治分歧的背景下，人类承担不起人工智能走向“逐底竞争”的后果。我们需要设立护栏，以便建立信任，确保人工智能安全、透明、负责任，并以人类尊严为中心。

江天骄分析，目前，主要AI大国的监管路径呈现显著差异。美国政策存在摆动效应，拜登政府时期强调建立监管联盟，而特朗普阵营则主张“轻监管”以优先鼓励技术创新；欧盟采取“严法”路径，通过《人工智能法案》实施严苛的风险分级与禁止性规定；中国则坚持“发展与安全并重”，通过备案机制、内容标识、溯源要求、“清朗”专项行动等灵活措施，实现场景化治理与技术创新发展的平衡。

谷歌、Anthropic、OpenAI和英伟达等美国多家科技企业负责人近日在白宫签署了名为《白宫超级智能协议：前沿责任联合承诺》的文件。根据这份协议，所有训练、部署前沿模型的企业，均需建立完善的内部流程与管控机制，确保技术按预设目标运行。该协议要求企业重点监测模型在网络安全、生物安全和化学威胁等领域的能力及相关风险，并防止模型以非预期方式入侵或访问技术系统。有美国媒体指出，上述承诺属自愿性质，并非具有法律约束力的监管规定。

今年9月，中国全国网络安全标准化技术委员会发布《人工智能安全治理框架3.0》。新的框架更加关注人工智能出现的新风险，设置“自主实施网络攻击威胁”专栏，还有专门的“智能体风险管理框架”。9月下旬，中美达成8点成果共识。其中包括双方同意建立中美人工智能对话，交流人工智能相关风险和惠益。

江天骄认为，面对AI技术的迅猛发展，加强安全监管与国际合作刻不容缓。一方面，全球多国需要探索成立全球公认的前沿模型第三方独立评估机构和互认机制，推动建立AI技术相关国际安全标准，为更多国家提供可信的AI模型安全检测评估服务；另一方面，国际社会仍需探索建立国际安全事故通报与披露机制，明确实验室及企业在发现风险时的通报义务、对象及内容，尤其是AI大国间的合作将产生重要示范效应。此外，国际社会还须完善AI模型分类分级标准，改变单纯依靠参数量级评判风险的做法，重点考察AI系统自主迭代优化、模型自主实施网络攻击、多智能体协同开展高危操作等实质性风险特征，为人工智能划定更加细致的安全“红线”。

中国“养蜂经”让肯尼亚妇女生活“甜”起来

“她们都像变了一个人，脸上的笑容更多了，也更加自信了。”阿霍利帕马·切拉加特·鲁托向记者介绍肯尼亚巴林戈郡莱尔科洛妇女小组成员时感慨道。

几年前，这些妇女大多没有稳定收入，家庭开支和孩子学费曾让她们头疼不已。自从鲁托从中国取来了“养蜂经”，她们不仅掌握了养蜂和采蜜技术，还开始经营自己的蜂蜜品牌。2025年，这个21人妇女小组共收获蜂蜜612公斤，销售收入约50万肯尼亚先令（约合2.5万元人民币）。

这一变化始于鲁托的一次中国之行。2017年，鲁托赴长沙参加由中国商务部主办、湖南省农科集团有限公司承办的发展中国家蜜蜂生产与粮油作物增产综合技术培训班。

在中国，鲁托第一次系统学习了蜂场建设、蜂群繁育和蜂箱管理，其中让她印象最深的是人工分蜂技术。“过去肯尼亚养蜂人通常只是把蜂箱放在那里，让蜜蜂自己进去。”她说，“在中国，我学会了如何主动将一个蜂群分成多群，实现规模化养殖。”

回到巴林戈后，鲁托把在中国学到的养蜂技术分享给当地妇女。“我希望她们能够获得更多能力和经济上的独立。”她说，在巴林戈，许多妇女承担着取水、做家务和照料家庭等工作，缺少稳定的收入来源。在她看来，养蜂占用日常时间较少，又能带来收入，适合当地妇女发展家庭副业。

在当地社区，养蜂长期被视为男性的活计。传统蜂箱通常放置在树顶，采蜜需要爬树且缺乏防护设备，妇女养蜂一度被视为禁忌。

鲁托开始向妇女推广朗氏蜂箱等现代蜂箱。她说，相比传统蜂箱，现代蜂箱降低了养蜂和采蜜的操作门槛，朗氏蜂箱可以安置在地面附近，配合防护服和专业工具，女性也能独立完成从检查到采收的全过程。

对莱尔科洛妇女小组来说，最初的尝试仍带着几分犹疑。2022年第一次收获蜂蜜时，妇女们仍习惯地请男性帮忙。“我们心里还保留着传统观念。”小组主席格蕾丝·切塞雷姆说。

第二次收获时，在鲁托的鼓励下，小组成员第一次亲手打开蜂箱。切塞雷姆仍记得自己第一次亲手收获蜂蜜的场景。“听着防护面罩下自己的呼吸声，小心翼翼地抽出第一块巢脾，金黄色的蜂蜜在阳光下清晰可见，周围的姐妹们发出了欢呼。”她说。

妇女们掌握养蜂技术后，在家庭中的角色悄然发生变化。“现在我们夫妻双方都可以分担家庭经济责任。”切塞雷姆说。如今，切塞雷姆拥有10个蜂箱并经营自己的家庭蜂场。



肯尼亚巴林戈郡莱尔科洛妇女小组成员观察蜂箱。 新华社记者 刘 琼摄

养蜂也提供了一种应对农业生产风险的新选择。莱尔科洛妇女小组秘书费思·基托尼说，当地部分地区今年旱情严重，玉米和豆类大面积减产。“如果没有养蜂的收入，我们的日子会非常艰难。养蜂成了我们应对气候不确定性的依靠。”

今年，妇女小组开始尝试拓展销售渠道，经营一家销售自有蜂蜜的商店，推出自己的品牌。这吸引了周边村庄的居民前来“取经”。与此同时，鲁托正将“蜂蜜增值加工”理念引入妇女小组，引导她们从销售原蜜向蜂产品加工延伸，提升附加值。“我计划明年培训她们进行其他蜂产品加工，比如制作身体乳、润唇膏、蜂蜡制品等。”

从一名赴华培训的学员，到如今巴林戈郡的一名农业推广员，鲁托一直尝试将所学技术在当地推广。迄今，她已将从中国学到的“养蜂经”传授给当地约84个养蜂小组。

一个个蜂箱不仅承载着收获的蜂蜜，更寄托着当地民众对美好生活的期待。正如基托尼所说，养蜂让她的生活“变得更加甜蜜”。 （据新华社电 记者由荟圆、王灏、刘琼）

在中亚国家感受“中文热”

城市轻轨站内，自助售票机屏幕亮起，中文与哈萨克文、俄文、英文四种文字并列呈现，指尖轻触，购票轻松完成；走出站台，网约车司机主动用略显生涩的中文问候“您好”；落座用餐，不少餐厅服务员会递上中文菜单……连日来，记者穿行中亚多国，目之所及，中文应用场景不断延展，一股“中文热”扑面而来。中文，正以肉眼可见的方式，融入中亚社会，走进当地人的日常生活。

夜幕降临，哈萨克斯坦最大城市阿拉木图喧嚣未歇。阿里·法拉比哈萨克国立大学孔子学院的教室里，琅琅跟读声与轻快问答声交织回荡。

26岁的阿依木·诺鲁孜巴耶娃供职于当地一家物流企业，长期对接中方客商的经历，让她看到中文蕴藏的机遇。今年9月，她报名学中文。“希望明年能去中国留学，学成归来继续在哈中合作领域工作。”她说。

这所孔子学院由兰州大学与阿里·法拉比哈萨克国立大学合作共建。学院中方院长邓红说：“近年来当地中文学习氛围持续升温，学员年龄从7岁到70岁不等。”学院是汉语水平考试全球考点之一，年均服务考生2000余人次，报考人数逐年上升。

相似的情景出现在塔吉克斯坦。塔吉克斯坦民族大学孔子学院中方院长孙静说：“学院每年招生两期，每期招收1000余人，但每期报名者都超过1500人。有的家长长途跋涉送孩子来学习中文。”

中亚民众学习中文的渠道日趋多元。记者走访发现，除孔子学院外，民间中文培训、中文家教市场正蓬勃兴起。在阿斯塔纳，记者在生活服务平台检索“中文课程”，找到136家中文培训机构；其他人口较多的城市也有数十至上百家同类机构。



吉尔吉斯斯坦贾拉拉巴德国立大学孔子学院，学生在中文课上学习。 新华社记者 江宥林摄

傍晚时分，当记者踏进阿拉木图“老师”教育中心时，这家民间中文培训机构教室里人坐得满满当当。儿童、在校学生、在职人员同堂学习，课堂气氛非常热烈。中心负责人塔季扬娜·考克诺娃说：“哈中经贸、物流、科技合作持续深化，主动学习中文的当地人逐年增多。”

家教是学习中文的新潮流。27岁的萝扎在阿里·法拉比哈萨克国立大学孔子学院完成学业后成了一名中文家教老师，目前正带着15名青少年学习中文。“中国文化很迷人，中文为我们打开了一扇观察中国的窗户。”萝扎说。

中亚兴起“中文热”并非偶然，而是经贸联通、文明互鉴、政策便利多重因素叠加的结果，中文正由课堂语言，变为中亚青年逐梦未来的“机遇密码”。

就读于贾拉拉巴德国立大学的吉尔吉斯斯坦青年拜卡扎克·苏丹别科夫学习中文两年有余。自今年8月20日起，吉尔吉斯斯坦公民可适用240小时过境免签政策和海南30天入境免签政策来华，“得知这个消息我特别开心，去海南看海的愿望越来越近了”。

“家人都全力支持我学中文。”17岁的穆斯利玛正在乌兹别克斯坦塔什干国立东方大学孔子学院学习，2025年在陕西西安研学时，她被古城风物深深打动，从此对中国文化产生浓厚兴趣。“我很想去中国多看看。”她说。

哈萨克斯坦中国研究中心主任古丽娜尔·沙伊梅尔格诺娃表示，中国在数字经济等高新产业领域的领先优势，使得人工智能、金融科技等领域的双语人才越发抢手，带动中文学习需求不断增长。

在中亚这片古老的土地上，语言承载着文明相遇的温度，让中国与中亚的合作故事在每一次诵读、交谈间，变得鲜活可感、温润动人。

（据新华社电 记者贺占军、黄思路、江宥林）

环球掠影

长春大冬会火种在意大利成功采集

近日，在意大利都灵，长春2027年第33届世界大学生冬季运动会火种成功采集。

图①：火种灯抵达都灵理工大学后，艺术家与大冬会吉祥物在现场表演。

图②：火种盆在火种采集仪式上被点燃。

新华社记者 王楷焱摄