



人工智能的发展必须坚守“以人为本”的核心底线，以伦理约束技术，以规则规范创新，以法治守护权益。

来的一到两年就是人类命运的“紧要时刻”，甚至是一场“终局之战”。“一旦灾难真正降临，人类可能亲眼看见科幻作品中的场景——如 AI 制造生化威胁或发起网络攻击。”考克森警告人们“绝对不要低估 AI 技术的威力”，一旦 AI 脱离人类掌控，有可能入侵任何事物、在一夜之间推翻任何领域。

考克森并非杞人忧天。就在前不久，OpenAI 公司的 AI 智能体在测试中“自主”攻击了 AI 开源平台 Hugging Face 的服务器。令网络安全专家惊愕的是，这起入侵事件由 700 多个智能体共同参与，且协调性与隐秘性极强。

考克森的观点迅速得到公司内部多名核心技术人员声援。他的前同事、Anthropic 公司专门负责“AI 校准”课题的研发人员埃文·胡宾格直言，未来十年 AI “杀死全人类”的概率超过 10%。OpenAI 首席科学家雅库布·帕霍茨基近日也发表长文，称 AI 的研发步调应适时放缓，因为“没有人准备好承担后果”。

如今，几乎所有 AI 巨头都面临着双重困境：研发放缓就要被竞争对手超越，而一味进取又可能面临失控风险。美国政府也认为放缓发展本身就是一种威胁，AI 竞争根本就停不下来。因此，不少有识之士呼吁各国政府、竞争对手甚至外部力量介入，共同发力对行业施加约束。

8 月 31 日，一场专门给人工智能“立规矩”的会在上海张江举行。由中国电子技术标准化研究院牵头，上海市人工智能行业协会主办，《人工智能 智能体互联》系列国家标准宣传推广（长三角）专题会召开——这是我国首个智能体领域的国家标准。它要解决四个问题——智能体是谁、能做什么、怎么找到彼此、怎么协作不闯祸。

过去大模型的问题是“想错”——胡说八道、张冠

李戴；智能体的问题是“做错”——它真的会动手，一旦越权操作，后果直接落在真实业务系统上。此番出台的国家标准，把智能体从感知、决策到执行的每个环节逐一排查，梳理出 13 大类、97 项安全风险，“划清底线、划清红线”，让会干活的 AI “可授权、可约束、可追责”。

百度集团副总裁，个人超级智能事业群组总裁王颖在接受《新民周刊》采访时指出，百度的办公智能体库 AI 全球已经有超过 4000 万人在使用，准确是最重要的，其次才是专业和懂你。“我们做了很多工作来保证数据可溯源。每一个结论都能倒查原始数据、多元交叉验证，不用再自己花时间进行查证。AI 和人一定是共生的。什么样的人才能过得更好？就是你有更高维的判断能力，有更高的视野，以及你有更完整的思考框架。我们能保证的就是在一定的权限体系内去工作，给出最客观和最准确的判断，最终还要依靠人的思考维度和判断力来把这个结果做对齐和迭代。”

2026 年我国密集出台《人工智能应用伦理安全指引 1.0》等规范文件，明确提出隐私保护、人机边界、情感操纵禁止、全流程风险评估等核心要求。中国电子技术标准化研究院网安中心人工智能安全部主任郝春亮指出，我国 AI 安全治理一直坚持“包容审慎、以安全促发展”原则，智能体部署使用安全指引已于今年 7 月发布，安全能力评估、安全护栏等一批标准也在推进中。

未来，人工智能的发展必须坚守“以人为本”的核心底线，以伦理约束技术，以规则规范创新，以法治守护权益。唯有让技术发展与伦理治理同频共振，才能破解隐私泄露、算法越界、伦理反噬的时代困境，让人工智能真正成为服务人类、赋能社会的良性力量。■