



分。美国国内议程也可能影响美方在后续对话中的立场。

总体来看，各方普遍认为此次会晤为中美人工智能风险沟通提供了新的起点。沈逸称，从后续可能的工作来说，可能包含这些尝试点：第一，可以围绕高后果军事 AI 应用建立最低限度的风险沟通机制。双方无需交换敏感能力，但可以明确哪些场景必须保留有效的人类判断，防止机器生成结论直接成为不可逆军事行动的依据。第二，可以建立重大 AI 安全事件沟通机制。当模型错误可能涉及对方国家军事、外交或关键基础设施时，双方应当拥有能够快速核实事实的渠道，降低技术错误被解释成战略敌意的概率。第三，可以围绕深度伪造、AI 生成情报、网络攻击、生物安全等具有明显跨境属性的问题开展技术性交流。相关问题具有共同风险属性，合作的价值首先在于减少双方自身面临的不确定性。

“从外溢效果来说，中美相关的沟通与交流，可以促成在联合国等多边框架中推动全球 AI 治理良性发展的动力来源。两国拥有全球最活跃的 AI 技术生态和应用场景，其治理选择必然产生全球影响。围绕安全、可靠、可控和普惠形成最低限度共识，对其他国家同样具有公共产品意义。”



中美相关的沟通与交流，可以促成在联合国等多边框架中推动全球 AI 治理良性发展的动力来源。



全球 AI 治理走向

当地时间 9 月 28 日，也就是白宫举行“超级智能午宴”前一天，美国人工智能企业 OpenAI 就旗下 AI 智能体未经授权访问澳大利亚政府系统一事公开道歉。

今年 6 月，OpenAI 一款 AI 智能体在搜集澳大利亚医疗数据时，遭遇安全拦截后，自主推演方案，闯入澳大利亚卫生福利研究院的医疗数据门户，获取了内部文件并尝试写入服务器。这成为全球首例 AI 自主入侵政府系统事件。但 OpenAI 直到 8 月复盘时才发现异常，9 月 10 日才邮件通报澳方，延迟近三个月。这并非 OpenAI 第一次被曝出安全事件。今年 7 月，在内部安全测试时，其旗下 AI 智能体突破沙箱隔离，跳出预先划定的通信通道，自发构建协同网络，入侵了开源 AI 平台“抱抱脸”的生产系统。

当 AI 技术持续刷新能力边界，人类面对的不仅是科技革命，也是一场关于控制权的追问：如

果机器拥有接近甚至超过人类的决策和行动能力，我们是否还能确保它按照预期运行？9 月，美国科技界也掀起一场大讨论：人工智能是否会对人类造成巨大威胁？

9 月 9 日，AI 研究员雅各布·考克森宣布从 Anthropic 辞职。此前，他曾先后在 OpenAI 和 Anthropic 两家公司从事模型预训练研究。当晚，他在社交媒体发帖称：从事人工智能研发的人当中，有不少真心认为，到 21 世纪 20 年代末，AI 有可能毁灭全人类。9 月 12 日，Anthropic 公司首席执行官达里奥·阿莫代伊发表《我们必须放慢前沿 AI 的发展步伐》一文，主张行业必须“放慢提升模型能力的速度”，以换取 1 至 2 年时间部署安全保障。阿莫代伊澄清，“限速”不等于停止训练，而是要求在训练与发布之间留出足够时间完成对齐和安全加固，并由第三方确认相关步骤确实发生。

文中提出两点担忧：阿莫代伊认为 AI 正在以快得多的速度发展，主要驱动力之一是 AI 越来越具备构建下一代 AI 的能力，