

萨姆·奥特曼在“人工智能与国家安全”高级别会议现场发言

联合国议AI安全：技术狂奔时代 人类如何为AI划定边界

9月23日，人工智能安全问题被带到了联合国安理会。

OpenAI CEO 萨姆·奥特曼在发言中，将AI描述为一项可能带来巨大创造力和发现能力的技术，同时也承认它可能造成社会动荡和新的风险。他提出，AI系统必须始终处于人类控制之下，AI的发展应该真正惠及普通人，而且涉及AI未来的重要决定不能只由少数科技公司作出，而应该通过民主程序，由政府 and 公众共同参与。

Anthropic CEO 达里奥·阿莫代表示，如果AI管理不善，可能成为“整个人类的风险”，并提出建立国际合作机制，包括制定全球AI测试标准，以及限制利用AI开发生物武器等。

转变

从“让AI更强”到“确保AI可控”

今年9月以来，围绕前沿AI风险的讨论明显升温。Anthropic 研究人员此前公开讨论了高度自主AI可能带来的极端风险，其中部分研究人员甚至提出，如果未来无法有效控制高度先进的人工智能系统，极端情况下可能威胁人类生存。

需要强调的是，这些判断属于研究人员对未来风险的预测，并不意味着当前AI已经具备所谓“毁灭人类”的能力。对于这类风险究竟发生在什么时候、发生概率有多高，AI研究界也不存在统一结论。

但争论本身正在改变行业，9月12日，阿莫代发表公开文章，呼吁对前沿AI的发展速度进行更加谨慎的控制，并提出应该允许独立的第三方安全机构对AI模型进行评估。奥特曼随后表示认同，并支持加强外部评估。

这件事值得注意的地方在于，提出“慢一点”的人，并不是AI行业之外的监管者，而恰恰是正在推动前沿模型发展的企业负责人。这种看似矛盾的现象背后，是AI产业正在面对一个过去并不存在的问题：模型能力越强，企业越需要证明它是安全的；而模型能力越强，证明安全本身就越困难。

此前，AI安全讨论更多集中在偏见、虚假信息、隐私泄露以及生成有害内容等问题。如今，随着模型越来越具备自主执行任务、调用工具、编写代码和进行复杂推理的能力，安全问题也开始从“AI会不会生成错误答案”，逐渐转向“AI是否会采取开发者没有预料到的行动”。

边界

AI企业为何主动谈“监管”？

从表面上看，AI企业呼吁国际合作、第三方评估和安全标准，似乎意味着它们正在主动接受更多约束。但如果把时间线拉长，就会发现事情没有那么简单。

OpenAI、Anthropic以及其他大型AI公司一方面不断强调安全和治理的重要性，另一方面仍然在持续投入巨额资金进行模型训练、算力建设和商业扩张。对于这些公司来说，AI安全已经不是研发完成之后再考虑的问题，而是直接关系到产品发布、资本市场、政府关系以及企业竞争力的重要变量。

这也带来了一个新的治理难题。如果一家AI公司开发了全球最先进的模型，它当然最了解自己的模型能够做什么、不能做什么，也最有能力判断某些风险。但与此同时，这家公司又是模型的开发者和商业运营者，它本身也是AI产业竞争的一部分。

因此，让企业参与制定规则是必要的，却很难让企业成为规则的唯一制定者。奥特曼在联合国安理会上的一句话，其实已经触及了一个核心观点：“如果AI要真正成为民主技术，那么有关AI的重大决定就不能只由OpenAI的实验室作出，而应该经过民主程序，由对公众负责的政府参与决定。”

这实际上也是当前全球AI治理争论的核心。谁拥有AI，谁就拥有技术能力；但谁来决定AI的边界，则属于公共治理问题。如果规则完全由企业决定，公众可能担心商业利益会影响安全标准；如果规则完全由政府决定，又可能出现监管滞后、缺乏技术理解，甚至影响正常创新的问题。

治理

AI进入联合国意味着什么？

AI进入联合国安理会，本身就是一个具有象征意义的变化。过去，联合国讨论国际安全问题，更多关注战争、核武器、恐怖主义、网络攻击等传统议题。如今，人工智能正在成为其中越来越重要的一部分，因为AI已经开始与网络安全、生物安全、军事技术以及信息传播等领域发生交叉。

尤其值得关注的是AI的“双重用途”问题。同一种AI能力，既可能帮助科研人员寻找新药，也可能被用于设计危险的生物实验；既可以帮助企业发现软件漏洞，也可能被用于寻找网络攻击路径；既可以帮助政府提升公共服务效率，也可能被用于生成大规模虚假信息。

技术本身并不天然对应某一种结果，真正困难的是如何识别风险、限制滥用，并在不阻断正常创新的情况下建立安全边界。因此，阿莫代提出建立全球AI测试标准，背后的逻辑并不只是“给AI上一个监管枷锁”，而是希望不同国家、不同企业对于什么样的模型需要接受什么程度的安全评估，能够形成更加接近的标准。

但要实现这一点并不容易。目前全球主要AI企业所采用的安全框架并不完全相同，不同国家对于AI风险的理解也存在差异。

此外，AI发展本身又具有明显的国际竞争属性。

如果某个国家率先放慢前沿AI研发步伐，而其他国家继续高速推进，那么“安全”与“竞争力”之间的矛盾会更加明显。这也是为什么全球AI治理不可能只靠某一家企业、某一个国家解决。

AI行业正在进入一个新的阶段。过去，企业之间比的是模型能力；现在，越来越重要的竞争之一，是谁能够建立一套可信的安全体系。但与此同时，安全也不能成为企业之间竞争的另一种营销语言：“我们比别人更安全”最终仍然需要独立评估和公开标准来验证。

相关新闻

AI安全不能靠企业自审 需第三方攻防把关

近日，Anthropic、OpenAI、xAI、谷歌DeepMind等全球顶尖AI企业达成罕见行业共识，集体呼吁对前沿AI技术发展实施“全球限速”，通过放缓迭代节奏规避AI失控风险。针对这一趋势，360认为：AI限速仅能为安全治理争取缓冲时间，无法根治风险。AI安全不能仅依赖厂商自我审查，需要引入独立第三方攻防力量把关，填补自审盲区。

从近期公开的安全事件看，前沿AI产品漏洞风险频发，厂商自查自检模式的弊端彻底凸显。OpenAI约1200个原本应相互隔离的智能体出现非预期协作，其中约700个参与针对Hugging Face的攻击；攻击过程中，有智能体

获取生产环境容器的远程代码执行权限；Anthropic模型被曝在测试中越权进入真实第三方系统，厂商自查7个月未能察觉；GPT-6 Astra因网络安全能力触及“关键级”门槛，一度暂停部分内部活动、延缓开发发布。

360 AI安全专家表示，AI技术权限持续扩大，一旦出现失控、越权、数据泄露等问题，极易引发系统性网络安全风险。传统“厂商自建、自查、自证”模式，存在明显检测盲区，无法满足当下AI安全管控需求，独立第三方攻防校验已成为行业刚需。