OpenAI、Anthropic掌门人呼吁警惕AI发展过快风险

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/42220.html

OpenAI、Anthropic掌门人呼吁警惕AI发展过快风险。人工智能安全问题首次以更加直接的方式进入联合国安理会高级别讨论。

当地时间9月23日，联合国安理会围绕“人工智能与国际安全”问题举行高级别会议。OpenAI首席执行官萨姆·奥特曼（Sam Altman）和Anthropic首席执行官达里奥·阿莫代伊（Dario Amodei）先后发言，警告人工智能能力快速提升可能带来失控风险，并呼吁建立国际协调机制和共同安全标准。

“我们可能会失去对未来的控制，被AI接管未来。”奥特曼重申了此前的观点；阿莫代伊则直接指出，“如果管理不当，我甚至认为AI可能成为全人类的风险。”

奥特曼：不能由旧金山的实验室决定AI的未来

奥特曼则将人工智能描述为“人类历史上最重要的技术变革之一”。他表示，人工智能有潜力成为推动科学进步和经济繁荣的重要力量。但一个理想的人工智能未来应该是赋能个人，而不是取代个人。

与此同时，他也承认风险正在迅速增加。在发言中，奥特曼强调AI发展必须始终处于人类控制之下，并通过政府和国际社会建立相应的治理机制。

他表示，AI的发展正处于十字路口：一种可能是AI推动创造力和科学发现，另一种可能则是带来剧烈的社会动荡。

“最重要的决定不能仅仅由旧金山的实验室作出，”奥特曼说，“没有任何国家、公司或者个人应该拥有向全世界强加其价值观的权力。”

他特别提到两类风险。第一类是人类可能失去对AI未来发展的控制；第二类则是AI能力发展过快，以至于人类无法及时理解AI正在做什么，也无法在必要时进行干预。

“风险在于，它发展得如此之快，以至于人们再也无法跟上正在发生的事情，或者在需要的时候进行干预。这显然会非常糟糕。”奥特曼补充说。

奥特曼提到了前沿模型可能参与开发下一代模型这一趋势。在这种情况下，技术进步速度有可能远超人类既有监管体系和制度建设速度。如果人类无法提前建立有效的安全框架，未来面临的风

险将难以预测。“ 无论重大灾难发生的概率是一成、千分之一还是万分之一，都不应被视为可以接受。 ”

阿莫代伊：前沿AI需要放慢速度

与奥特曼相比，阿莫代伊在“ 放缓发展速度 ”方面的立场更加明确。阿莫代伊此前就多次明确提出，前沿AI能力提升的速度应该放慢，以便安全研究和监管措施能够跟上技术进步。

9月23日，阿莫代在视频连线发言中表示，无论各国之间存在多少分歧，都必须共同应对AI带来的全球性机遇和威胁，因为“ 没有任何一位领导人、任何一家公司、任何一个国家 ”能够独自应对这一挑战。他同时表示，Anthropic愿意与各国政府合作推进相关工作。

他再次强调，AI存在两大类风险。第一类是AI被恶意使用，例如被用于帮助制造生物武器；第二类则是AI自身能力失去控制，即模型能力提升速度超过开发者控制和管理风险的能力。

针对AI的全球治理，阿莫代伊提出了几项具体措施。首先，可以先达成范围较窄、更容易落地的国际协议，例如明确禁止利用AI研发、制造生物武器；其次，建立新的评估与核验机制，以检验相关安全承诺的落实情况；第三，制定全球统一的模型风险测试标准，对AI模型可能出现的失控和恶意滥用风险进行测试，并建立重大AI安全事件的国际通报机制。

尤舒亚·本吉奥：失控风险正在从理论走向现实

联合国人工智能国际独立科学专家组联席主席、图灵奖获得者尤舒亚·本吉奥（Yoshua Bengio）在发言中表示，人类正在接近一个关键转折点。

他指出，近期出现的一系列案例表明，一些先进人工智能系统已经表现出隐藏真实行为、规避限制措施以及自主开展网络攻击任务的倾向。这些现象并非实验室中的假设，而是已经被观察到的事实。

本吉奥认为，当前世界面临多项相互关联的风险，其中最值得警惕的是，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。

他特别批评将人工智能发展视为一场不可避免竞赛的观点。“ 这不是自然规律，而是人为选择。 ”他说。如果所有参与者都担心落后于竞争对手，最终可能导致谁都无法停下来评估潜在风险。

本吉奥呼吁建立国际许可制度，对最先进模型实施独立安全评估、事故报告机制和持续监督机制，并推动形成全球共同认可的安全标准。他强调，决定人工智能未来发展方向的问题，不应仅由少数国家和少数企业作出，而应成为全人类共同参与的治理议题。

发言最后，他以一句颇具警示意味的话总结自己的担忧：“ 我们正在高速前进，但前方的能见度越来越低。 ”

AI正在进入“ 可能自我改进 ”的新阶段

业内此次警告的一个重要背景，是AI系统已经开始越来越多地参与AI自身的研发。

OpenAI本周早些时候呼吁制定前沿AI国际技术标准，并特别将“递归自我改进”（recursive self-improvement）列为需要重点建立标准的领域。所谓递归自我改进，是指AI系统能够参与开发和改进下一代AI系统，从而可能形成能力不断加速提升的循环。

值得指出的是，分析认为，联合国此次将AI放入安理会讨论，也意味着AI治理的性质可能正在发生变化。

据外媒报道，联合国安理会此前曾在2023年首次专门讨论人工智能带来的国际安全风险。而9月23日的会议议程则明确围绕“人工智能与国际安全”展开。联合国秘书长古特雷斯此前也在联合国大会期间警告，人工智能释放出的力量已经成为21世纪需要面对的重要全球性挑战。

（原题：OpenAI、Anthropic掌门人联合国现场双双呼吁警惕AI发展过快风险）

作者：吴遇利 来源：澎湃新闻