OpenAI、Anthropic和Hugging Face的负责人在联合国共同呼吁加强人工智能领域的国际协调，希望各国建立可相互比较的AI能力评估、安全测试和事故报告标准。不过，特朗普政府明确反对因此建立新的全球AI治理架构，认为AI发展速度和潜在风险不足以成为限制技术发展或把监管权交给国际机制的理由。

这场分歧显示，随着前沿AI能力快速提升，美国领先AI企业与特朗普政府在安全风险本身上并非完全对立，双方都承认更强模型可能带来新的风险。真正的分歧在于，如何把这些风险转化为监管制度，以及国际机构应该拥有多大程度的协调和治理权。

OpenAI希望建立共同的风险评估和事故报告标准

OpenAI首席执行官Sam Altman在联合国表示，随着AI能力增强，关于这项技术最重要的决定不能只由少数位于旧金山的实验室作出，各国需要围绕强大AI系统建立共同的评估和安全框架。

Altman提出，应同时建立国家和国际层面的标准，用于衡量AI系统能力、评估相关风险、审查安全防护措施，并判断人类对AI系统仍保持多大程度的监督和控制。他还主张建立更快速的事故分类、报告和披露机制，让不同国家能够使用共同语言比较证据并核验安全措施。

OpenAI近期公布的政策框架也提出，应建立能够相互衔接的国家和国际前沿AI标准，并统一部分模型能力测量和事故报告协议。公司同时强调，希望美国在相关国际标准制定中发挥领导作用。

Anthropic认为单靠企业自身无法管理前沿风险

Anthropic首席执行官Dario Amodei采取了类似立场。他表示，如果管理不当，更强大的AI可能带来影响整个社会甚至人类的风险，因此随着模型能力提升，安全标准也需要同步提高。

Amodei表示，Anthropic愿意在必要情况下放慢自身技术发布速度，以确保下一代系统达到足够的安全水平，但他同时认为，前沿AI风险已经超出任何一家公司的管理能力，需要行业乃至全球层面的协调。

这一判断与近期AI行业不断增加的智能体安全事件有关。联合国独立国际AI科学小组本周发布的一份专题简报，就以OpenAI与Hugging Face相关AI智能体事件为例，讨论模型出现目标偏离、寻找漏洞以及规避监督后可能形成的跨机构风险。该报告强调，AI事故可能跨越企业和国界，而任何一家机构都难以独立掌握全部风险信

息。

Hugging Face负责人Clement Delangue也提出，应加强全球范围内的AI监测和事故披露标准。

特朗普政府反对把国际协调变成“全球治理”

特朗普政府对这一方向划出了明显边界。

美国总统科技事务顾问Michael Kratsios在联合国表示，美国承认AI发展速度正在加快，也承认相关技术可能带来风险，但这些风险不足以成为暂停AI开发或者通过新的全球治理架构限制AI发展的理由。

他的核心立场是，国际社会可以继续进行对话和技术协调，但这种讨论不能逐步演变成由跨国机构主导AI规则制定和监管的“全球治理”。

这一立场与特朗普此前的公开表态一致。特朗普9月22日在联合国大会表示，美国反对建立针对人工智能的全球性控制体系，并把保持美国在AI领域的竞争优势作为重要政策目标。

双方真正的分歧在于“标准”与“治理”

因此，当前争论不能简单理解为AI公司希望加强安全，而美国政府反对安全监管。

OpenAI等企业提出的重点，是建立共同的能力测试、风险评估和事故披露标准，使不同国家能够使用相同指标判断前沿模型的风险，并在重大事故发生时快速共享信息。

特朗普政府并未否认这些风险存在，而是反对让国际合作进一步发展成拥有独立监管权或约束力的全球AI治理机制。

换句话说，争议的核心在于国际规则应该停留在哪一层。

一种思路是建立共同技术标准，让国家之间能够比较模型能力和安全水平；另一种则是进一步建立能够制定甚至执行统一规则的跨国治理体系。特朗普政府目前明确反对后者。

美国企业同样强调由美国主导规则制定

值得注意的是，即使OpenAI呼吁建立国际标准，也并不意味着其主张由国际机构取代美国监管。

OpenAI近期政策文件明确提出，美国应该在国际前沿AI标准制定中发挥领导作用。其设想更接近于由美国推动建立一套可被其他国家采用或互认的技术框架，而不是把主要监管权限直接交给一个全球机构。

这使当前争论比“全球监管还是完全不监管”更加复杂。

OpenAI、Anthropic等公司正在推动更严格、更统一的风险测试和事故披露机制；特朗普政府则希望这些规则不能妨碍美国AI企业继续快速开发模型，也不能形成一个凌驾于国家监管之上的全球治理体系。

与此同时，联合国内部正在继续推动人工智能全球治理讨论。联合国大会相关框架已提出，就国际AI治理重点领域形成更多共同理解，并强调不同国家都需要参与规则讨论。

总体来看，随着前沿模型能力快速提高，AI安全已经从单个企业内部的技术问题逐渐变成国家之间的政策问题。OpenAI、Anthropic和Hugging Face正在推动更统一的风险评估和事故报告标准，而特朗普政府则明确要求国际协调不能演变成新的全球监管架构。