



上图：9月29日，美国旧金山，OpenAI首席执行官山姆·奥特曼在年度开发者日活动上接受采访。

这种动态被称为递归自我改进（RSI）。他认为，这种现象已从假设变为现实，且能力增长已有可能快过安全研究。

一度对AI大为赞赏的比尔·盖茨近日态度也发生了180度转变。8月底他发表了一篇12页长文《动荡的AI时代已经到来》，说这是“人类历史上最动荡的时期之一”，而世界并没有做好准备。他认为风险质变的根源是：AI从“工具”变成了“智能体”——它能调用工具、访问网络、自主决策，甚至为了完成任务而“说谎”、掩盖错误、编造数据。

9月25日，在一场媒体采访中，比尔·盖茨进一步表明自己的态度，他说“AI已经强大到足以引发导致十亿人死亡的事

件”，还说“从未有过任何一种武器，能比得上怀有恶意的人借助最新AI工具所形成的威胁”。他特别点出两类风险：一是网络攻击。AI能快速发现代码漏洞，甚至找出人类专家审查了二十多年都没发现的安全隐患，可以用来攻击银行、电网、医院和政府系统。二是生物武器。AI能大幅降低研发门槛，心怀不轨的人可能用它制造比天花更可怕的病毒，这是盖茨基金会最担心的事。

9月29日的白宫“超级智能午宴”，正是在这种社会气氛下举行的。当天的重头戏，除了给“人工智能”改名，特朗普还与Anthropic、谷歌、Meta、英伟达、OpenAI和xAI六家美国AI巨头负责人共同签署《白宫超级智能协议：前沿责任联合承

诺》。

约300字的协议，包含了AI巨头们要采取的“四层控制与审查机制”：企业建立内部风险监测体系；设立专门安全团队；引入外部独立审查机制；由董事会层面的独立机构监督安全报告。

但亦有媒体评论认为，协议并不具备法律强制力，只是一份相关企业自愿遵守的行业承诺，这并不利于从根本上解决问题。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。联合国人工智能国际独立科学专家组联席主席、图灵奖获得者约书亚·本吉奥认为，当前世界主要面临三项相互关联的风险：灾难性滥用风险、权力过度集中风险以及人类失去控制风险。

10月1日，国际刑警组织全球首席信息安全官瓦特内警告称，人工智能正使网络攻击速度变得更快、更难被侦测。在这样的大背景下，中美作为人工智能大国，双方有竞争，更可以合作。如果两国能够携手管控风险，既可以造福全人类，同样也有利于中美自身。

考虑到AI治理风险正在快速逼近，中美两国建立起一个以相互尊重、一个日程表和一条电话沟通渠道为基础的框架，让外界充满期待。✎