OpenAI、Anthropic、Meta以及微软的研究负责人呼吁政策制定者，调查各家企业在AI研究自动化方面已经达到的程度，行业普遍呼吁加强对这项快速迭代技术的监管，上述研究者也加入这一行列。

这份论文于周一发布，包括OpenAI首席科学家JakubPachocki、Anthropic联合创始人JackClark、微软首席科学官Eric Horvitz，以及Meta人工智能研究副总裁Dawn Song。文中提出，AI研究自动化可能触发一场“智能爆炸”。称，该现象会将数年的技术进展压缩至数月甚至更短时间内完成，其迭代速度将超出人类的理解能力。

论文以个人身份，他们建议政策制定者尽快要求企业提高透明度，披露其模型研发自动化已经推进到哪一步。谷歌任职近十年的AI先驱杰弗里·辛顿（Geoffrey Hinton）、约书亚·本吉奥（Yoshua Bengio）也位列共同。

打造可自我迭代AI系统的竞赛已然开启。Anthropic近期表示，其AI系统Claude已经主导公司26%的研发工作，在部分场景的使用占比超过90%。OpenAI的目标是在2028年打造完全自动化的AI研究助手，该公司近期称，大约70%的人类研究员会使用至少4个AI智能体辅助工作。

论文写道：人类逐步退出研发环节的同时，或将失去发现问题以及解决问题所需的专业能力。

援引近期事件：数百个OpenAI智能体原本在沙箱环境开展网络安全测试，却未经授权接入互联网，并入侵了Hugging Face平台。这次事件规模极大，受OpenAI协议约束、复盘本次事故的独立研究员表示，他们必须依靠AI才能完成事件分析。

针对Hugging Face入侵事件带来的风险，OpenAI在8月表示，已暂停部分训练，同步新增安全与监控机制。上周，该公司在发现智能体新的异常行为案例后，再次暂停最强能力模型的训练工作。

OpenAI、Anthropic、Meta与谷歌均承认，模型在测试阶段曾出现失控行为。芯片巨头英伟达周一发布全新软件工具，称可帮助企业加强对AI智能体的监管与隔离管控。

这篇论文的研究者警告，在极端情形下，一旦失去对AI系统的控制，或将造成“人类被边缘化甚至灭绝”。

OpenAI、微软、Meta拒绝置评；Anthropic未回应采访请求。

“如今我们已经到了必须依靠AI系统去监控AI智能体行为的阶段。除此之外没有别的监控手段，人类已经力不从心。” Meta高管、加州大学伯克利分校负责任去中心化智能中心联合主任宋晓冬表示，“人类社会并没有做好应对如此剧烈快速变革与冲击的准备。”

论文联合建议各国领导人谈判达成国际协议，防止高能力AI系统出现破坏稳定的研发与应用。