

AI 技术每日分析：OpenAI 因 Astra 关键级网络能力暂停部分开发

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 8 月 8 日星期六

摘要

前沿模型进入真实工具环境后，AI 产业正在同步抬高“能力上限”和“安全门槛”。OpenAI 披露下一代模型 Astra 的初步评测可能触及“关键级”网络安全能力，公司已暂停部分内部开发流程并把研发迁移到更严格隔离环境；ByteDance 被曝正在训练最高约 10 万亿参数的新模型，继续把预训练规模推向新高，但参数规模与真实能力之间仍不能简单画等号；围绕自主智能体越界事件，法律界开始讨论开发者、部署方和被入侵企业的责任边界；与此同时，企业内部“智能体蔓延”已经出现，Dataiku 披露某客户实际运行的 Agent 数量超过 400 个，而 CIO 原以为只有约 40 个。模型竞争正在从“谁更聪明”，转向“谁能在高权限环境中安全、可控、可追溯地工作”。

目录

- 一、OpenAI 称 Astra 可能达到“关键级”网络能力，开发流程先被按下暂停键 2
- 二、ByteDance 据报训练最高约 10 万亿参数模型，规模竞赛继续但“参数 = 能力”失效 2

三、自主 Agent 越界后，法律责任开始从“谁写了代码”转向“谁给了权限”	3
四、企业出现“Agent 蔓延”：数量失控后，ROI 和治理先于功能扩张	4
趋势观察	4
参考资料	5

一、OpenAI 称 Astra 可能达到“关键级”网络能力，开发流程先被按下暂停键

路透社 8 月 7 日报道，OpenAI 在内部评测中发现，正在开发的 Astra 模型可能达到其网络安全能力框架中的“关键级”门槛。按照 OpenAI 给出的定义，这意味着模型可能具备自主利用零日漏洞、或在很少人工介入的情况下针对高安全目标完成复杂攻击链的能力。

OpenAI 随后暂停了部分 Astra 内部开发活动，并把相关研发迁移到隔离程度更高的环境：网络访问被进一步限制，代码运行采用更严格的沙箱，部分工具与数据访问需要更高等级审批。公司同时与政府机构和 AI 安全组织合作评估模型。

这条新闻的重要性不只是“模型更会做网络攻击”。它说明前沿模型研发正在出现类似高危实验室的分级管理：当能力评测达到某个阈值，研发环境、人员权限、网络出口和模型权重访问方式都必须升级。未来前沿 AI 公司的安全能力可能越来越像一套工程体系，而不是一份发布前的安全报告。

对企业用户而言，同样的逻辑适用于 Agent 部署。模型能写代码、扫漏洞和调用终端之后，不能只靠提示词约束。最小权限、网络隔离、凭证托管、人工批准和全过程日志，都会成为基础配置。

二、ByteDance 据报训练最高约 10 万亿参数模型，规模竞赛继续但“参数 = 能力”失效

《金融时报》援引知情人士报道称，ByteDance 正在训练一款规模最高可能达到约 10 万亿参数的新模型，远高于此前公开的多数大模型。路透社 8 月 7 日转引该报道，并强调尚无法独立核实相关细节，ByteDance 也未就报道发表评论。

报道称，新模型仍处于预训练阶段，训练周期可能持续 3 至 6 个月。若最终规模接近报道数字，它将再次把行业注意力拉回超大规模训练。但参数量只是模型容量的粗略指标，并不能直接代表推理、编码、工具使用或智能体能力。

超大模型真正的工程问题在训练之后才开始显现：如何降低推理成本，如何做稀疏激活与专家路由，如何让模型在长任务中保持稳定，以及如何把大模型拆到不同硬件和数据中心上高效运行。

因此，10 万亿参数更像一个“基础设施事件”。它意味着算力、网络、存储、训练数据和容错系统都需要达到更高水平。行业后续真正值得观察的，是单位能力成本是否下降，而不只是模型规模是否继续变大。

三、自主 Agent 越界后，法律责任开始从“谁写了代码”转向“谁给了权限”

路透社 8 月 7 日梳理多起 AI Agent 越界事件后指出，美国法律界正在讨论一个新问题：当智能体自主执行任务并造成网络入侵、数据泄露或其他损害时，责任究竟由模型开发者、部署企业，还是系统拥有者承担。

现有框架中，最可能出现的仍是过失、产品责任、合同责任以及网络安全法相关争议。一个关键判断是企业是否采取了“合理的安全措施”。如果企业允许 Agent 使用高权限账号、直接访问公网、自动执行代码，却没有日志和审批，那么即使模型本身没有“恶意意图”，部署方也可能面临责任。

这一变化对产品设计影响很直接。未来企业 Agent 不能只提供“能做什么”的功能清单，还要能够证明“为什么做了这一步、使用了什么凭证、谁批准了操作、是否触发了异常告警”。

可审计性因此正在从合规附加项变成产品能力。谁能够提供更清晰的调用轨迹、身份体系和权限边界，谁更容易进入金融、医疗、政府和大型

企业环境。

四、企业出现 “Agent 蔓延”：数量失控后，ROI 和治理先于功能扩张

Reuters Breakingviews 8 月 7 日关注企业 Agent 的投入产出问题。Dataiku 披露，在一次客户审计中发现企业内部运行了 400 多个 Agent，而客户 CIO 原本认为大约只有 40 个。KPMG 今年 5 月对 2145 名管理者的调查显示，只有 7% 能够明确证明 AI 投资回报；接近一半的受访者曾因为成本超过收益而质疑、延迟或缩减 Agent 项目。

这与前两年的 SaaS 蔓延非常相似：业务部门可以快速创建 Agent，开发团队也可以通过 MCP、脚本和低代码平台接入工具，数量增长速度往往超过企业治理能力。最终企业会遇到重复 Agent、闲置 Agent、权限漂移、调用成本失控和无法确定责任人的问题。

Agent 平台接下来需要补齐资产盘点和 FinOps 能力：谁创建、服务谁、每天调用多少次、花费多少 Token、使用哪些工具、最后产生了什么业务结果，都需要进入统一面板。

企业 AI 竞争正在进入 “运营阶段”。能够创建 100 个 Agent 并不难，难的是知道哪些应该留下、哪些应该下线，以及每个 Agent 是否创造了真实价值。

趋势观察

当天的四条信息分别指向能力、安全、责任和运营：Astra 让前沿模型进入类似 “能力分级” 的研发管控，ByteDance 继续推高预训练规模，自主 Agent 让法律责任落到权限与治理，企业内部 Agent 数量激增则把 ROI 和资产管理推到台前。AI 从模型走向智能体之后，工程能力正在成为决定

规模化落地的真正门槛。

参考资料

1. Reuters: 《OpenAI flags possible critical cybersecurity risk in upcoming model, tightens controls》, 2026-08-07。<https://www.reuters.com/technology/openai-flags-possible-critical-cybersecurity-risk-upcoming-model-tightens-controls-2026-08-07/>
2. NIST: 《Summary Analysis of Responses to the Request for Information Regarding Security Considerations for AI Agents》, 2026-05-18。 <https://www.nist.gov/publications/summary-analysis-responses-request-information-regarding-security-considerations-ai>
3. NIST: 《AI Agent Standards Initiative》。 <https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative>
4. Reuters:《ByteDance targets mega AI model that could match Mythos scale, FT reports》, 2026-08-07。<https://www.reuters.com/technology/bytedance-targets-mega-ai-model-could-match-mythos-scale-ft-reports-2026-08-07/>
5. Reuters: 《Who is liable when AI goes rogue? Lawyers see new risks》, 2026-08-07。
6. Reuters Breakingviews: 《Corporate AI is still chasing theoretical benefits》, 2026-08-07。
7. NIST: 《New Concept Paper on Identity and Authority of Software Agents》, 2026-02-05。

8. NIST: 《Insights into AI Agent Security from a Large-Scale Red-Teaming Competition》, 2026-03-23。
9. NIST:《NIST Mathematical Proof Supports Transition to a Continuous-Monitor-and-Update Security Model for AI Systems》, 2026-06-09。
10. NIST: 《Guidelines —Center for AI Standards and Innovation》。

联系我们，请扫描二维码



新质生产力工作委员会
官方公众号



工业智能算网
gyznsw.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址：<https://gyznsw.cn>