NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/14/the-ai-ceos-found-the-brakes-sign-language-ai-should-pay-attention/.

人工智能公司高管们找到了刹车……手语人工智能应当引以为鉴

Heather Grizzle · September 14, 2026

Frontier AI CEOs are now asking for testing, audits and incident reporting. Sign language AI should hold itself to the same brakes.

目录

1. 前沿正在呼吁踩刹车
2. 治理债务是真实存在的
3. 前沿正在学习的东西，手语人工智能现在就可以学到
4. "We tested it" is not governance
5. 不要等待属于你自己的“放慢脚步”时刻
6. 独立保障并非对创新的攻击
7. 抢在前面行动的机会仍然存在
8. 警示

前沿正在呼吁踩刹车

多年来，人工智能领域的主导叙事一直是速度。构建更强大的模型。发布智能体。击败竞争对手。筹集资金。扩展基准测试。把产品推向市场。治理可以稍后再补上。

“稍后”似乎已经到来。

On September 9, OpenAI chief global affairs officer Chris Lehane published a call for mandatory, capability-based national AI safety regulation covering testing standards, independent assessments, cybersecurity protections and incident reporting, alongside the company's endorsement of four California bills, including one establishing infrastructure for independent safety assessments and another setting standards for

AI auditors.[^1] Three days later, Anthropic CEO Dario Amodei went considerably further. In an essay arguing that the industry must deliberately pace frontier development, he proposed embedding third-party evaluators inside frontier labs with near-employee access so they could independently examine safety measures, incidents and model behavior.

随后，发生了一件不同寻常的事。

Within hours, OpenAI CEO Sam Altman publicly agreed that the frontier needed pacing, and told Fortune days later that his company would not pursue a public offering this year because it had too much safety work left to do. Elon Musk, who had pointedly declined to join earlier industry calls for restraint, quoted the essay approvingly. Demis Hassabis, now chair of Google DeepMind and chief scientist of Alphabet, wrote that the essay pointed toward the right path forward. Companies locked in one of the most consequential technology races in history converged, inside a single day, on the proposition that capability could no longer be allowed to outrun control. Axios ran the moment under a headline that needed no elaboration: "AI's most powerful CEOs hit the brakes."

Today, September 14, Microsoft added another piece. Microsoft AI CEO Mustafa Suleyman released a draft code of conduct that he describes as a constitution of sorts for the company's future models. The draft would require Microsoft's systems to communicate clearly with humans, respect human boundaries, remain controllable, accept correction and shutdown, and treat behavior that violates the code as a model failure rather than an acceptable expression of autonomous judgment. Microsoft assembled it after expert consultations and has opened it to roughly six weeks of public feedback.[^3] Suleyman called it a warning shot.

这一系列事件应当成为每一家开发手语人工智能的公司的必读材料。并非因为一个手语数字人即将接管互联网，而是因为这正是“治理债务”的模样。

治理债务是真实存在的

科技公司理解技术债务。开发得足够快，推迟足够多的架构决策，最终这些走的捷径就会变得代价高昂。

治理债务的运作方式相同，它是通过一系列单独看来都合理的决策不断累积起来的。一家公司在明确系统失败时谁应负责之前就发布上线。它在确定何为错误之前就发布准确

性声明。它在未询问基准测试是否代表真正会使用该产品的人群之前就构建了基准测试。它进行内部验证，因为独立评估太不方便。它加入人工监督，却未说明该监督人员应注意什么、何时必须介入，或他们是否具备识别失败的专业能力。它在重要的产品决策已经做出之后才任命一个咨询小组。它把一个能生成手势的系统称为“无障碍”。

上述每一个选择都在制造治理债务。而与技术债务不同的是，其成本并非只由开发者承担。用户承担了它。采购方承担了它。公共机构承担了它。在手语技术领域，聋人承担了它。

前沿正在学习的东西，手语人工智能现在就可以学到

当前的人工智能讨论中存在一个非同寻常的事实：那些拥有全球最庞大的工程团队、研究预算和安全机构的公司，正在要求更多的外部审查。Anthropic正在提议嵌入式第三方评估。OpenAI正在支持强制性联邦监管和独立评估。微软正试图正式确立人类控制的边界。竞争对手的高管们正公开讨论协调一致的克制，而这在不久前还是难以想象的。

这里的教训并不是公司应该等待政府制定出完美的标准。恰恰相反，最接近这项技术的机构正在承认，内部的信心并不足以构成充分证据。

这对手语人工智能而言至关重要，因为这一领域仍然足够年轻，可以选择一条不同的发展顺序。治理不必是事后的修补工作，它可以成为架构的一部分。

"We tested it" is not governance

手语人工智能存在一个尤为危险的验证难题。一个系统可以生成看起来像手语的东西，却仍然辜负依赖它的人。一段翻译可以流畅却是错的。一个数字人可以在视觉上很精致，却引入了语言上的扭曲。一个基准测试可以报告一个亮眼的数字，却代表着一个狭窄的数据集、一个受限的打手语环境，或一个与实际部署几乎毫无相似之处的任务。一项研究可以包含人工验证，却依赖一个远远太小的样本，去支撑日后附加在其上的商业宣称。而一次产品演示可以效果出色，却并未证明该系统适用于医疗、教育、就业、紧急通讯或政府服务。

这些不仅仅是机器学习方面的问题，更是保障性的问题。谁测试了它，谁选择了测试内容？哪些打手语者被纳入代表，哪些没有？测量了什么，又排除了什么？什么算作实质性错误？评估了哪些部署场景？谁有权停止部署？当性能下降时会发生什么，届时又由

谁来审查事故？而对于围绕聋人沟通而构建的技术，聋人的话语权究竟在哪个环节真正进入决策？

这些问题在合同已经签署、营销宣传已经发布、客户已被告知技术已经就绪之后，会变得难以回答得多。

不要等待属于你自己的“放慢脚步”时刻

There is an understandable temptation for smaller AI companies to look at the frontier governance debate and file it under somebody else's problem. Different technology, different scale, different risks. All correct. But risk magnitude is not the relevant comparison. Governance timing is.

前沿行业在异常激烈的竞争压力下急速前进，如今其领导者们正面临关于独立评估、强制测试、事故报告、人类控制、政府监管，以及该行业是否值得信任、能否自行判断其系统是否足够安全等问题。手语人工智能开发者可以实时旁观这一切的展开，然后决定是否要重蹈覆辙。

今天开发手语技术的公司，应当已经能够回答以下五个问题。

1. 什么样的证据会促使我们不部署自己的产品？
2. 公司之外，谁有权质疑我们的宣称？
3. 除了提供反馈或训练数据之外，聋人还拥有怎样的决策权？
4. 即便客户愿意付费，我们认为哪些用途仍然是对我们技术不恰当的使用？
5. 当部署后的证据与我们当初用来推销该系统的证据相矛盾时，会发生什么？

如果这些问题的答案并不存在，那么这家公司拥有的就不仅仅是尚未完成的治理文档，它拥有的是治理债务。

独立保障并非对创新的攻击

这正是手语人工智能市场需要迅速成熟的地方。独立审查并非对该技术的敌意，而是表明这项技术足够重要，值得被审视。

飞机制造商不是靠邀请买家欣赏飞机来建立信心的。医疗器械制造商也不是靠发布用户证言来确立安全性的。财务审计不允许管理层的热情代替证据。随着人工智能进入无障

碍领域，同样的区分依然适用。开发者创造系统。治理确立围绕它的规则。评估检验证据。保障则追问这些证据是否足以支撑所提出的宣称。这是四种不同的职能，当同一家公司同时掌控这四种职能时，购买者理应提出质疑。

The frontier industry is beginning to acknowledge exactly that problem. Anthropic's proposal for embedded outside evaluators is significant precisely because it moves independent scrutiny closer to the development process rather than treating evaluation as a ceremonial checkpoint immediately before release.

手语人工智能应当走得更远。独立的技术评估是必要的，但语言的合法性、部署场景、无障碍影响以及聋人治理，都无法简化为工程测试。一个模型可以完全按设计运行，却仍然不适合完成它被出售用来完成的工作。

抢在前面行动的机会仍然存在

这正是手语人工智能公司不应错过的部分。它们被给予了一个异常宝贵的警示，无需先亲身经历一场治理危机。

要做的工作是具体的。在营销文案将局限性美化之前，先记录下这些局限。将演示表现与部署证据区分开来。在客户提出要求之前，先主动使用独立评估者。在第一起严重事故发生之前，先建立事故处理流程。明确哪些场景下产品将不会被出售。赋予聋人专家真正的决策权，而非仅仅是咨询性的在场。让验证声明足够具体，使另一方能够审视并质疑它们。保存证据，以便购买者日后能够确认，某项部署决策在做出当时为何是合理的。

这就是“可被保障”所应有的样子。而且，这样做也比日后重建信任要便宜得多。

警示

世界上最强大的AI公司正在发现，治理不可能永远落后于能力。它们中的一些领导者现在正在呼吁踩刹车。另一些则在呼吁立法。还有一些正在起草章程、外部评估结构和新的控制机制。它们掌握着大多数无障碍技术公司永远无法企及的资源，即便如此，它们也在发现，一旦技术已经全速前进，治理会变得多么困难。

Sign language AI companies should study that carefully. Do not wait until a failed deployment, a procurement challenge, a regulator, a lawsuit or the Deaf community forces the governance conversation. Do not build the evidence after making the claim.

Do not invite independent scrutiny only after trust has been lost. And do not confuse being first to market with being ready for the responsibility that comes with entering someone else's language, communication and access.

前沿AI现在正试图找到自己的刹车。手语AI仍有机会在加速之前就将刹车装好。

NOVARA CONSULTING GROUP

Novara Consulting Group 提供AI保障、无障碍治理与独立评估服务。

FOOTNOTES:
Chris Lehane, "The AI policy window is open. We need to act.," OpenAI, September 9, 2026; Reuters, "OpenAI pushes for mandatory national AI safety requirements," September 9, 2026.
Axios, "AI's most powerful CEOs hit the brakes," September 13, 2026.
Jeffrey Dastin, "Microsoft drafts code of conduct to keep its AI under human control," Reuters, September 14, 2026.

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

SUPPORT NCG

NCG does this work without vendor money, so our findings stay ours. If it has been useful to you, you can support NCG's independent work here. NCG is a for-profit company, not a charity. Contributions are not tax-deductible and buy no service or influence.

引用方式

Grizzle, H. M. (2026, September 14). 人工智能公司高管们找到了刹车……手语人工智能应当引以为鉴. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003722