# AI能够翻译。但谁来为风险翻译担责?

Heather M. Grizzle · August 17, 2026

*AI can translate, but who translates the risk? Explore the AI accountability gap, responsible AI governance, procurement standards, and sign language AI oversight.*

**目录**

1. 什么是AI问责缺口?
2. 能力、可靠性与问责是三种不同的证据
3. 为何披露法律无法弥合这一缺口
4. 验证不对称:潜藏其下的结构性问题
5. 无障碍法早已部分回答了这个问题,而我们正在遗忘它
6. 为何供应商演示不能作为治理记录
7. 机构在部署前应当要求什么
8. The next stage of AI is verification
9. 常见问题
10. 资料来源

**每一次*AI*部署都会把风险转嫁给某个人。在大多数情况下,风险落在最没有能力发现错误、也最没有权力提出质疑的人身上。**

Heather M. Grizzle Novara Consulting Group 2026年8月17日

Earlier this month I wrote about Google DeepMind's sign language translation research and argued that a demonstration is not the same thing as evidence. The response told me something useful. Almost nobody disagreed that the distinction matters. Many people asked the obvious follow-up question instead: if a demonstration is not enough, who is supposed to decide when a system is good enough to deploy, and who answers for it when it is not?

这个问题并非手语所独有。它是当前AI治理领域尚未解决的核心问题,而且它有一个值得直接使用的名字。姑且称之为问责缺口。

## 什么是AI问责缺口?

The AI accountability gap is the distance between what a system is capable of doing and what any identifiable party is answerable for when it does that thing badly. It opens whenever an organization deploys an AI system faster than it builds the means to verify the system′s output, hear from the people the output affects, and correct or stop the deployment when the output is wrong.

这一缺口首先并不是一个技术问题。模型的失败方式是可以预见的,工程师们通常也会坦率地承认这些失败模式。这个缺口是一个制度问题。各机构正在采购其无法评估输出的系统,将其部署到错误会带来后果的场景中,却没有保留任何切实可行的机制来查明系统是否正常运作。

## 能力、可靠性与问责是三种不同的证据

大多数采购讨论把这三者当作一回事。它们并不是一回事,而将它们区分开来,是机构采购方能做出的最有价值的一步。

**Capability evidence** shows that a system can perform a task under conditions the vendor selected. Demonstrations, benchmark scores, and pilot footage are capability evidence. They answer the question "is this possible."

**Reliability evidence** shows how the system performs across the range of people, settings, and conditions the buyer will actually encounter, including the ones the vendor did not choose. Disaggregated performance data, independent testing, and error rates from live deployment are reliability evidence. They answer the question "how often is this wrong, and for whom."

**Accountability evidence** shows what happens when the system fails. It names who monitors performance, who receives complaints, who has authority to suspend the deployment, what the affected person is owed, and how any of that is enforced. It answers the question "who is responsible."

Vendors compete on capability evidence because it is the cheapest to produce and the most persuasive to watch. Reliability evidence is expensive and unflattering.

Accountability evidence is not really the vendor's to produce at all, because it describes the buyer's own governance, and most buyers have not built it.

一个机构如果把能力证据当作已经解决了另外两个问题,那它其实并没有做出采购决策。它做出的是一个假设,并为此付出了代价。

## 为何披露法律无法弥合这一缺口

目前对AI风险最引人注目的监管回应是透明度。根据欧盟AI法案(European Union Artificial Intelligence Act)第50条,自2026年8月2日起生效的义务要求提供者以机器可读形式标记合成内容,并要求部署者在涉及公共利益的事项上披露深度伪造内容和某些AI生成的文本,违反者最高可被处以1500万欧元或全球年营业总额百分之三的罚款。欧盟委员会为已经投放市场的生成式系统给予了直至2026年12月2日的宽限期。

这项工作意义重大,我也支持它。但同样值得精确指出它究竟做了什么。第50条是一套溯源制度。它告诉一个人,某台机器曾参与其中。它并不会告诉这个人,这台机器是否做对了。

For a great deal of synthetic media, provenance is the whole question. If the concern is a fabricated video of a public figure, knowing the video is synthetic resolves it. But for AI systems that mediate communication, provide information, or influence a decision about a person, provenance is the easy half. A label reading "this translation was generated by AI" leaves the recipient exactly where they started, which is unable to tell whether the content is accurate.

The liability half of the European approach has moved in the opposite direction. The Commission's proposed AI Liability Directive, which would have eased the burden of proof for people harmed by AI systems, was formally withdrawn in October 2025. Transparency advanced. Redress did not.

这一模式在美国州一级重复出现。科罗拉多州于2024年通过了首部全面的州级AI法律,随后将其废除并以SB 26-189取而代之,该法于2026年5月14日签署,其对开发者和部署者的实质性义务自2027年1月1日起适用。替代法案取消了强制性风险管理项目、算法影响评估,以及针对算法歧视的独立合理注意义务。保留下来的是使用前告知、在不利决定作出后三十天内以通俗语言进行披露、三年的记录保存,以及请求有意义的人工复审的权利。

我并不是说这些保留下来的义务毫无价值。告知和记录保存都是实实在在的东西。但告知只是让你知道使用了某个系统。记录只是让你知道它做了什么。二者都不能证明输出是准确的,也都没有在输出有误时明确任何责任归属。过去两年监管走向的方向,是让人们知道AI参与其中,却回避了让任何人为其产出负责。

## 验证不对称:潜藏其下的结构性问题

这里是我认为被忽视的部分,也正是手语案例不再只是一个小众例子、而开始成为对一种普遍状况最清晰说明的地方。

在大多数AI部署中,最有能力发现错误的人没有权力采取行动,而有权采取行动的一方却没有能力发现错误。我把这称为验证不对称,它正是产生问责缺口的机制。

想想一个健听机构在部署自动手语系统时看到了什么。它看到系统产出了结果。工作人员看到一个虚拟形象在打手语,或者字幕出现在屏幕上。从这个角度看,交流似乎已经完成。

Now consider what the Deaf person experiences. They receive output they cannot check against the source, delivered by a system they did not select, in a setting where the institution has already concluded that access was provided. If the rendering drops a negation, flattens a conditional, misplaces spatial reference, or omits the non-manual markers that carry grammatical meaning, the person may not know. If they do notice something is wrong, they are in the position of contesting the institution′s own conclusion that communication succeeded.

这不是一个技术层面的抱怨。这是权力的分配。机构握有判定是否实现了无障碍访问的权威,却缺乏评估这一点所需的语言能力。聋人拥有这种语言能力,却没有相应的权威。如果存在错误,它便无处浮现。

一旦你看清了这种结构,就会发现它几乎无处不在。病人无法评估AI辅助的分诊建议。求职者无法审查将其筛除的模型。福利申领人无法审计资格认定。任何一位收到机器翻译的法律指示的人,若要核实翻译内容,恰恰需要那份翻译本应提供、却尚未具备的语言能力。在每一种情形中,受影响的一方都在承担一种他们无法衡量的风险,而部署机构则记录下一笔已完成的交易。

使手语成为一个格外尖锐的案例的原因,在于其内容所牵涉的利害关系。机构场景中的语言承载着同意、证词、诊断、指示与法律效力。误译不仅仅是一次体验上的降级,而是一份有缺陷的同意书、一份不准确的病史,或是庭审记录中被曲解的陈述。

## 无障碍法早已部分回答了这个问题,而我们正在遗忘它

无障碍实践中其实已经存在一个行之有效的答案,这使得当下对它的忽视更加难以自圆其说。

Under Department of Justice regulation at 28 CFR 35.160, a public entity must give primary consideration to the auxiliary aid or service requested by the person with a disability. The rule for private public accommodations at 28 CFR 36.303 is weaker, requiring consultation while leaving the final choice with the provider, and that difference is itself instructive about where verification authority tends to erode. The National Association of the Deaf's minimum standards for video remote interpreting in medical settings state the underlying principle directly: a deaf or hard of hearing individual knows best which auxiliary aid or service will achieve effective communication. Those same standards go further and require the video interpreter to tell the provider to terminate the remote session and obtain an on-site interpreter when the interpreter determines the remote arrangement is not producing effective communication.

如果把这条规则当作一种治理设计而非无障碍规则来看,它的构建异常出色。它把验证判断的权力交到真正有能力做出判断的人手中。它赋予现场的专业人员在使用过程中中止部署的权威。它把有效性视为一种需要结合具体情境判断的事物,而非仅凭设备的存在就想当然地认定。

Automated systems tend to strip both features out. There is no qualified interpreter positioned to call a halt, because the interpreter is what the system replaced. And the affected person's judgment about whether communication worked is quietly demoted from a legal standard to customer feedback.

世界聋人联合会(World Federation of the Deaf)与世界手语传译员协会(World Association of Sign Language Interpreters)早在2018年就为手语虚拟形象划定了一条界限。他们的立场是:虚拟形象不适用于现场的、复杂的或重要信息的传达,而预先录制的静态内容在聋人参与了内容审定、且不需要现场互动的情况下,或许可以被接受。这

一声明已有八年历史。此后技术已有长足进步。但它所划定的部署边界,至今没有被任何更严格的东西所取代,它仍然是目前可用的最清晰的界线。

## 为何供应商演示不能作为治理记录

演示的设计目的就是要有说服力。这是它的功能,本身并无不妥。问题在于,在缺乏独立评估的情况下,演示便默认成为采购决策的证据基础,因为它是现场唯一的文件。

一次演示所确立的,是在特定选定条件下的可能性。它无法确立系统在采购方所服务的全部人群中的表现、在受控条件之外的行为、在高风险场景中的安全性,或是出错时的问责机制。这些是部署的属性,而不是模型的属性。

联邦政策原则上已经认识到这一点。管理与预算办公室(OMB,Office of Management and Budget)于2025年4月发布的指南将高影响AI定义为其输出构成决策或行动主要依据、并对权利或安全产生法律性、实质性、约束性或重大影响的系统,该备忘录列举的领域包括公民权利、受教育机会、住房、就业、政府福利与服务以及医疗健康。对于这些用途,它要求进行部署前测试、在部署前及此后定期进行AI影响评估、实施充分的人工监督,以及为受影响个人提供及时的人工复审与申诉机会。无法达到这些最低要求的机构必须安全地停止使用。配套的采购备忘录要求合同条款须赋予机构定期监测和评估该系统性能、风险与有效性的能力。

无论人们对细节有何看法,这份指南的整体结构是正确的。它把问责视为一种应在合同中明确规定、在上线前经过测试、上线后持续监控、并可通过中止来强制执行的事项。这是一层验证机制。大多数在联邦采购体系之外采购AI的机构,并不具备与之相当的东西。

关于假定能力会按计划到来并自行弥合这一缺口,我想补充一点提醒。2026年4月,司法部将《美国残疾人法》(ADA,Americans with Disabilities Act)第二章网络无障碍规则的合规截止日期,针对每一类实体各推迟了大约一年,大型公共实体的截止日期推迟至2027年4月26日,规模较小的实体则推迟至2028年4月26日。司法部给出的理由之一是,它此前高估了受监管实体在人员配备和技术方面的能力,并明确指出,诸如生成式AI这样的先进技术,尚不能可靠地实现大规模自动化的无障碍内容修复。这是一个联邦监管机构记录在案的事实——在无障碍这一领域,AI能力的宣称超出了实际交付水平。在把任何供应商的时间表当作治理方案之前,这一点值得深思。

## 机构在部署前应当要求什么

以下这些问题并不高深。它们是任何有能力的采购方,对任何会带来后果的系统都会提出的问题,而且是可以得到回答的。

是谁评估了这个系统,他们是否独立于供应商和采购方?

测试中纳入了哪些人群、方言、地区变体和场景,每一项的结果分别是什么,而不是汇总结果?

这个系统是用什么数据训练的,基于何种许可,又征得了那些语言被采集者何种同意?

在与我们类似的条件下,其实测错误率是多少,当条件变差时,这一比率又会如何变化?

Who in our organization has the authority to suspend this deployment, and what triggers that authority?

How does an affected person report that the output was wrong, who receives that report, and what changes as a result?

What are we contractually entitled to know about performance after go-live, and what is our remedy if performance falls?

An organization that cannot answer these questions has not evaluated a system. It has accepted one. Novara Consulting Group built the Sign Language Access Trust Index to make these questions answerable in a form procurement can actually use, with defined evidence standards, indicator-level scoring, and a distinction between what a vendor asserts and what an independent party can confirm. The instrument is specific to sign language AI. The underlying discipline is not.

## The next stage of AI is verification

The public conversation about AI has been organized around capability for three years. Can it write, can it reason, can it see, can it act, can it interpret human communication. Those questions are being answered quickly, and often in the affirmative.

The question that determines whether any of it is trustworthy is different. It is whether the institutions deploying these systems can tell when the systems are wrong, and whether the people affected by them have any standing to say so.

Right now, in most deployments, the answer to both is no. The risk has not been translated. It has been transferred to whoever is least able to see it.

Closing that gap is not a matter of waiting for better models. It requires evaluation independent of the vendor, evidence standards that survive contact with procurement, authority to stop a deployment located with someone in the room, and a channel through which the affected person's judgment counts for something. None of that is technically hard. It is institutionally inconvenient, which is a different problem and a more tractable one.

---

## 常见问题

**What is the AI accountability gap?** The AI accountability gap is the distance between what an AI system is capable of doing and what any identifiable party is answerable for when it does that thing badly. It appears when organizations deploy AI faster than they build the ability to verify its output, receive complaints from affected people, and suspend the deployment when it fails.

**Is AI transparency the same as AI accountability?** No. Transparency rules, including Article 50 of the EU AI Act, generally establish provenance, meaning they tell people that content was generated or manipulated by AI. Accountability establishes who is responsible for the accuracy of that content and what the affected person is owed when it is wrong. A system can be fully labelled and still be unaccountable.

**Who is legally responsible when an AI system causes harm?** In most jurisdictions this remains unsettled. The European Commission withdrew its proposed AI Liability Directive in October 2025, and several United States state frameworks have narrowed toward notice and record-keeping obligations rather than substantive duties of care. In practice, responsibility is determined by contract terms, sector-specific law such as disability and civil rights statutes, and general liability doctrine rather than by AI-specific rules.

**Why is sign language AI a useful test case for AI governance?** Because it makes verification asymmetry visible. The institution deploying the system usually cannot evaluate the output, and the Deaf person who can evaluate it usually has no

authority to challenge the institution's conclusion that access was provided. The same structure operates in medical, hiring, and benefits AI, but is harder to observe.

**What should an organization require before deploying an AI system that affects people?** Independent evaluation rather than vendor demonstration, performance data disaggregated by population and setting, disclosed training data provenance, a named person with authority to suspend the deployment, a functioning complaint channel for affected people, and contractual rights to monitor performance after launch.

**What is the SLAT Index?** The Sign Language Access Trust Index is Novara Consulting Group's evaluation instrument for sign language AI systems. It sets defined evidence standards across governance, linguistic, transparency, accessibility, and deployment domains, scores systems at the indicator level, and distinguishes vendor assertion from independently confirmable fact so that procurement decisions rest on evidence rather than demonstration.

---

*Heather M. Grizzle is the founder of Novara Consulting Group, which advises institutions on the governance and procurement of AI accessibility technology. She is the author of the SLAT Index Evaluation Standard.*

*Related reading: The Demo Is Not the Evidence: What Google DeepMind's Sign Language AI Must Prove Next*

---

## 资料来源


- European Commission, Quick Facts: Transparency rules for AI systems (Article 50, applicable 2 August 2026). https://digital-strategy.ec.europa.eu/en/factpages/quick-facts-transparency-rules-ai-systems
- European Parliament Legislative Train Schedule, AI Liability Directive (withdrawal, October 2025). https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-ai-liability-directive
- U.S. Department of Justice, Extension of Compliance Dates for Nondiscrimination on the Basis of Disability: Accessibility of Web Information and Services of State and Local Government Entities, 20 April 2026. https://www.federalregister.gov/documents/

2026/04/20/2026-07663/extension-of-compliance-dates-for-nondiscrimination-on-the-basis-of-disability-accessibility-of-web

- Office of Management and Budget, M-25-21, Accelerating Federal Use of AI through Innovation, Governance, and Public Trust (April 2025). https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf
- Office of Management and Budget, M-25-22, Driving Efficient Acquisition of Artificial Intelligence in Government (April 2025). https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-22-Driving-Efficient-Acquisition-of-Artificial-Intelligence-in-Government.pdf
- Colorado SB 26-189, repealing and reenacting the SB 24-205 framework, signed 14 May 2026, substantive obligations applying from 1 January 2027. https://leg.colorado.gov/bills/sb26-189
- 28 CFR 35.160, Communications (Title II, primary consideration). https://www.ecfr.gov/current/title-28/chapter-I/part-35/subpart-E/section-35.160
- 28 CFR 36.303, Auxiliary aids and services (Title III, consultation standard). https://www.ecfr.gov/current/title-28/chapter-I/part-36/subpart-C/section-36.303
- National Association of the Deaf, Minimum Standards for Video Remote Interpreting Services in Medical Settings. https://nad.org/minimum-standards-for-video-remote-interpreting-services-in-medical-settings/
- World Federation of the Deaf and World Association of Sign Language Interpreters, Statement on Use of Signing Avatars, 14 March 2018, updated 14 April 2018. https://wasli.org/wp-content/uploads/2023/07/WFD-and-WASLI-Statement-on-Avatar-FINAL-14032018-Updated-14042018-1.pdf

---

## 引用方式

Grizzle, H. M. (2026, August 17). AI能够翻译。但谁来为风险翻译担责?. Novara Consulting Group. https://www.novaracg.com/zh/2026/08/17/ai-can-translate-but-who-translates-the-risk/