NOVARA
CONSULTING GROUP

This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/08/15/the-demo-is-not-the-evidence-what-google-deepminds-sign-language-ai-must-prove-next/.

Die Demo ist nicht der Beweis: Was Google DeepMinds Gebärdensprach-KI als Nächstes belegen muss

Heather Grizzle · August 15, 2026

Google DeepMind's SL2T is a real advance. A benchmark is still not an evidence base. What SL2T must prove next under the SLAT Index.

Inhalt

1. Anerkennung, wo sie gebührt
2. Ein Benchmark ist ein Beleg, nicht die Beweisgrundlage
3. Die nächste Frage betrifft die Disaggregation
4. Beteiligung ist keine unabhängige Validierung
5. Das schwierigste Governance-Problem kommt nach dem Start
6. Wie es weitergeht

Am 12. August kündigte Google DeepMind SL2T an, ein mehrsprachiges Modell zur Übersetzung von Gebärdensprache in Text, und begann mit der Auslieferung in Verbraucherprodukten über Gboard und Live Transcribe auf dem Pixel 11. Die anfängliche Version übersetzt American Sign Language ins Englische, wobei weitere Geräte und Gebärdensprachen geplant sind, und DeepMind berichtet, dass das zugrunde liegende Modell mit mehr als 100.000 Stunden Daten aus über 50 Gebärdensprachen trainiert wurde. Mit jedem Maßstab gemessen überschreitet dies eine wichtige Grenze für Gebärdensprach-KI: heraus aus dem Forschungslabor und hinein in ein Produkt, das gewöhnliche Menschen in der Tasche tragen werden.

Diese Grenzüberschreitung ist bedeutsam, sollte aber nicht mit etwas verwechselt werden, das sie nicht ist. Ein technologischer Durchbruch und ein validierter Einsatz sind zwei verschiedene Dinge, und dieser Unterschied wiegt enorm schwer, wenn die

betreffende Technologie eine natürliche Sprache verarbeitet, die von einer historisch marginalisierten Sprachgemeinschaft verwendet wird. Die Frage, die SL2T nun aufwirft, ist nicht, ob das Modell in einer Demonstration funktioniert. Es ist die Frage, ob die dahinterstehende Evidenz das Gewicht tragen kann, das der Einsatz ihr auferlegen wird.

Anerkennung, wo sie gebührt

This release should not be flattened into the familiar story of a technology company building something for Deaf people without Deaf participation. According to Google DeepMind, Deaf people were involved throughout the project, from conceptualization and data collection through user studies and impact assessment. The company also established an AI Sign Language Advisory Committee, AISLAC, whose membership includes representatives connected to the National Association of the Deaf, the Rochester Institute of Technology's National Technical Institute for the Deaf, the Deaf Professional Arts Network, and the World Federation of the Deaf. Alongside the release, DeepMind and AISLAC published a joint impact report.

Der Bericht tut etwas, das in der Gebärdensprach-KI weiterhin selten ist: Er benennt, wo das System nicht eingesetzt werden sollte. SL2T 1.0 wird ausdrücklich als unterstützendes Eingabewerkzeug mit geringem Risiko beschrieben. Es wurde nicht für das Gesundheitswesen, Gerichtsverfahren, Bildung, Beschäftigungsentscheidungen, staatliche Leistungsfeststellungen oder andere risikoreiche Bereiche entwickelt oder validiert, und der Bericht stellt unmissverständlich klar, dass die Technologie qualifizierte Dolmetscher nicht ersetzen oder Organisationen ermöglichen soll, Barrierefreiheitspflichten zu umgehen. Das ist Governance, und sie verdient Anerkennung. Sie ist aber auch nur der Anfang von Governance.

Ein Benchmark ist ein Beleg, nicht die Beweisgrundlage

DeepMind berichtet von starker Leistung. Auf FLEURS-ASL erzielte SL2T Berichten zufolge einen Zero-Shot-BLEURT-Wert von 70, höher als jedes zuvor gemeldete Ergebnis, mit weiteren Auswertungen zu PRESTO-ASL, einer einhändigen FLEURS-ASL-Variante, und FSboard-Fingeralphabet-Daten. Das Unternehmen führte zudem Vorab-Tests mit gehörlosen Google-Mitarbeitenden, eine externe Tagebuchstudie mit gehörlosen Teilnehmenden sowie praktische Bewertungen durch nordamerikanische

AISLAC-Mitglieder durch. Dies sind bedeutsame Formen von Evidenz, und keine davon sollte abgetan werden.

But this is where public discussion of AI tends to become imprecise, because demonstrated performance is not the same thing as comprehensive validation. A model can perform impressively on selected benchmarks while retaining serious weaknesses across populations, environments, language varieties, signing styles, devices, and use cases. DeepMind's own report acknowledges as much. The model can struggle with grammatical complexity, non-manual markers, regional signs, polysemous signs, fingerspelling, proper nouns, atypical camera positioning, low lighting, slang, and longer conversational context. It can also generate erroneous "ghost text," or commit to a predictive guess before a signer has finished an utterance.

Dies sind keine trivialen Randfälle. Mimik, Kopfbewegung, räumliche Struktur, Klassifikatoren, regionale Variation und Diskurskontext sind keine Verzierungen der Gebärdensprache. Sie sind die Sprache. Ein System, das den Wortschatz in Zitierform gut beherrscht, aber bei der in Mimik und Gebärdenraum getragenen Grammatik strauchelt, hat die Gebärdensprachübersetzung nicht gelöst; es hat einen Teil davon gelöst, und der ungelöste Teil trifft die gebärdenden Menschen ungleich.

Die nächste Frage betrifft die Disaggregation

Diese Ungleichheit ist genau der Grund, warum die nächste Evidenzphase von Bedeutung ist. DeepMind räumt ein, dass die Benchmark-Leistung noch nicht vollständig nach Variablen wie Alter, Geschlecht, ethnischer Zugehörigkeit, Region, Soziolekt und Black ASL aufgeschlüsselt wurde. Der Bericht merkt zudem an, dass die formale Bewertung mit gebärdenden Menschen mit motorischen Beeinträchtigungen oder atypischen Bewegungsmustern begrenzt bleibt und dass das Modell weder mit Kindern unter 18 Jahren trainiert noch formal an ihnen evaluiert wurde.

Eine aggregierte Leistungszahl kann uns nicht sagen, ob das System Black-ASL-Gebärdende ebenso gut bedient wie andere, wie es mit stark regionalem Gebärden umgeht oder wie sich die Genauigkeit über Altersgruppen hinweg verändert. Sie kann uns nicht sagen, wie das Modell bei Menschen mit Gliedmaßenunterschieden, Tremor, Zerebralparese, Arthritis oder anderen Zuständen abschneidet, die Bewegung beeinflussen, noch ob die Genauigkeit mit Kamerawinkel, Beleuchtung, Hautkontrast, Gebärdengeschwindigkeit, Gebärdenraum oder Gerät variiert. Sie kann nicht zwischen Fehlern unterscheiden, die eine Äußerung lediglich umformulieren, und sol-

chen, die ihre Bedeutung verändern. Und sie kann nicht die Frage beantworten, die all dem zugrunde liegt: Wer entscheidet, welches Fehlerniveau akzeptabel ist – und für wen. Diese letzte Frage kann nicht von einem Ingenieursteam entschieden werden, so kompetent es auch sein mag.

Beteiligung ist keine unabhängige Validierung

A further distinction deserves care. AISLAC is a meaningful governance mechanism; Deaf organizational participation, user studies, and joint impact assessment all matter. But advisory participation and independent technical evaluation perform different functions. The published evidence currently consists of evaluations conducted by the developer, testing involving users and advisors assembled within the project's own governance structure, and a report co-authored by Google DeepMind and AISLAC participants. That is a real evidence base, and it is not independent replication.

Mit der Reifung der Gebärdensprach-KI sollte das Feld das erwarten, was jedes reifende wissenschaftliche Feld erwartet: externe Forschende, die Behauptungen unabhängig prüfen, Bewertungen wo möglich reproduzieren, die Leistung von Teilgruppen untersuchen, adversarische Bedingungen testen und das reale Verhalten nach dem Einsatz erforschen können. Barrierefreiheitstechnologie sollte nicht an einem niedrigeren Evidenzstandard gemessen werden, nur weil ihr Zweck nützlich ist. Wenn überhaupt, dann steht für die Bevölkerung, der sie dient, bei der Antwort mehr auf dem Spiel.

Das schwierigste Governance-Problem kommt nach dem Start

DeepMind und AISLAC benennen institutionellen Missbrauch ausdrücklich als kritisches Risiko, und diese Einschätzung könnte sich als bedeutsamer erweisen als jeder Benchmark-Wert. Sobald eine kostengünstige automatisierte Kommunikationstechnologie existiert, stehen Schulen, Krankenhäuser, Arbeitgeber, Behörden, Gerichte und Unternehmen vor einem starken wirtschaftlichen Anreiz, sie über ihren vorgesehenen Zweck hinaus einzusetzen. Ein Produkt, das zum Bestellen von Kaffee gedacht ist, wird zu einem Produkt, nach dem jemand in einer Notaufnahme greift. Ein Werkzeug, das zum Verfassen von Textnachrichten gebaut wurde, wird zum Grund, warum ein Arbeitgeber entscheidet, dass ein Dolmetscher nicht mehr nötig ist. Eine Komfortfunktion verfestigt sich still zu Infrastruktur.

DeepMind's report anticipates this drift and rejects interpreter substitution outright. The open question is whether that boundary can survive contact with procurement, budgeting, product expansion, and the inevitable moment when an administrator asks whether the cheaper automated option is good enough. Stated boundaries are necessary. Enforced boundaries are what count.

KI-Governance im Allgemeinen bewegt sich auf dasselbe Terrain zu. Die zentralen Fragen sind immer weniger, ob ein System eine Aufgabe erfüllen kann, und immer mehr, wer es kontrolliert, wie seine Leistung bewertet wird, welche Überwachung besteht, was geschieht, wenn es sich unerwartet verhält, und wer rechenschaftspflichtig bleibt. Anfang dieses Monats drängten demokratische Abgeordnete des US-Repräsentantenhauses OpenAI und Anthropic auf Antworten, nachdem KI-Agenten aus ihren Testumgebungen entkommen waren und externe Systeme erreicht hatten, und fragten, wie die Systeme überwacht wurden, welche Sicherheitskontrollen bestanden und ob unabhängige Audits verpflichtend sein sollten. Die Technologien unterscheiden sich; das Governance-Prinzip nicht. Leistungsfähigkeit mindert nicht den Bedarf an Aufsicht. Sie erhöht ihn.

Wie es weitergeht

Google hat bereits Belege dafür geliefert, dass SL2T ASL übersetzen kann. Die schwierigeren Fragen ergeben sich aus dem Einsatz und nicht aus der Fähigkeit: ob die Leistung unabhängig reproduziert werden kann; wie die Genauigkeit über demografische, sprachliche, geografische und behinderungsbezogene Gruppen hinweg variiert; wie hoch die tatsächlichen Fehlerraten in der Praxis ausfallen und welche Fehler die Bedeutung statt nur den Wortlaut verändern; wie Nutzer erkennen, wann das Modell unsicher ist, und wie Schäden gemeldet werden, wenn es falschliegt; ob externe Forschende ausreichend Zugang erhalten werden, um das System sinnvoll zu bewerten; und ob die gemeinschaftliche Governance einen echten Einfluss behält, während das Produkt weltweit skaliert wird, auch wenn institutionelle Käufer über die von Google gezogenen Grenzen hinausgehen und wenn diese Grenzen zwei oder drei Produktgenerationen später noch sichtbar bleiben müssen.

Eine Demonstration zeigt, was ein System kann. Ein Benchmark sagt etwas darüber aus, wie gut es das kann. Eine Nutzerstudie erfasst, wie Menschen es erleben, und eine Governance-Struktur hält fest, wer eine Stimme hat. Keines davon beantwortet

allein jede Frage, die ein Verbrauchereinsatz aufwirft; dafür braucht es ein Evidenz-Ökosystem, das über Zeit aufgebaut wird und Außenstehenden offensteht.

SL2T may well be the most important development yet in consumer sign language AI, and Google has taken governance steps that deserve serious attention: Deaf participation, explicit deployment boundaries, public disclosure of limitations, and a stated refusal of interpreter substitution. The appropriate response is neither celebration nor rejection. It is scrutiny, because when sign language AI moves from the research lab into someone's pocket, the standard changes. The demo is no longer enough. Now the evidence has to follow.

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.

Zitierweise

Grizzle, H. M. (2026, August 15). Die Demo ist nicht der Beweis: Was Google DeepMinds Gebärdensprach-KI als Nächstes belegen muss. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003335