This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/09/14/microsofts-humanist-ai-code-of-conduct-governance-aspiration-and-procurement-reality/.

Aspirazione di governance e realtà dell'approvvigionamento: una lettura del Codice di Condotta per l'IA Umanista di Microsoft

Heather Grizzle · September 14, 2026

Microsoft's Humanist AI Code of Conduct is a north star, not a contract. Where it is precise, where it is not, and what buyers should ask.

Contenuti

1. Ciò che il quadro normativo coglie nel segno
2. Dove la precisione operativa viene meno
3. Il divario settoriale per i sistemi di lingua dei segni
4. Questioni strutturali che il documento lascia aperte
5. Che cosa significa questo per gli acquisti
6. Conclusione

Approfondimenti di Novara Consulting Group

Heather M. Grizzle

Microsoft AI ha pubblicato il proprio Code of Conduct per i modelli MAI il 14 settembre 2026, aprendo una consultazione pubblica di sei settimane sul testo. Il documento delinea i comportamenti, i valori e le salvaguardie previsti per la famiglia di modelli prodotta da Microsoft AI, organizzati attorno a una filosofia progettuale che l'azienda definisce Humanist AI. Due fatti sul suo stato contano più di qualunque elemento del suo contenuto. Primo, Microsoft afferma che il documento non viene attualmente utilizzato per addestrare i propri modelli e che una versione rivista, elaborata sulla base

della consultazione, è destinata a guidare lo sviluppo a partire dal 2027. Secondo, l'azienda descrive il documento come una stella polare piuttosto che una descrizione del comportamento attuale dei modelli, e riconosce uno scarto tra ciò che esso specifica e ciò che i suoi modelli fanno attualmente. Chiunque legga questo testo come uno strumento di conformità, una garanzia o una base per una rassicurazione istituzionale lo sta leggendo per qualcosa che esso non pretende di essere.

Questa distinzione non è un tecnicismo. Gli addetti agli acquisti ricevono abitualmente documenti di governance dei fornitori come prova che un sistema è sicuro da implementare, e tali documenti sono abitualmente scritti in un registro che invita a questo fraintendimento. Microsoft è stata insolitamente schietta su questo punto, e tale schiettezza merita di essere riconosciuta prima che inizino le critiche. L'analisi che segue tratta il documento per quello che dichiara di essere: una bozza di filosofia di governance sottoposta a consultazione, offerta esattamente per il tipo di scrutinio che questo articolo applica.

Ciò che il quadro normativo coglie nel segno

La struttura di autorità è la parte più solida del documento. Microsoft definisce una catena di comando in cui il Code of Conduct si colloca al di sopra delle politiche degli operatori, le quali a loro volta si collocano al di sopra delle preferenze degli utenti, e designa una categoria di vincoli assoluti che né gli operatori né gli utenti possono configurare per aggirarli. La stratificazione è coerente, e i requisiti di controllo umano che l'accompagnano, i quali riguardano interruzione, arresto, limiti di ambito e integrità delle tracce di azione, sono sufficientemente specifici da essere riconoscibili come requisiti ingegneristici piuttosto che come dichiarazioni di intenti. Il divieto per i modelli di oscurare il proprio ragionamento o di comunicare in forme che gli esseri umani non possono verificare è un autentico impegno di governance, poiché preserva le condizioni che rendono possibile qualsiasi supervisione a valle.

Il trattamento del fallimento della sicurezza come bidirezionale è il secondo punto di forza. Il documento individua sia la cautela insufficiente, in cui un modello favorisce un danno, sia la cautela eccessiva, in cui rifiuta richieste legittime o richiede conferme superflue, e afferma che il secondo tipo di fallimento si verificherà probabilmente più spesso, anche se il primo è più grave. La maggior parte dei quadri di sicurezza dei fornitori ottimizza per il rifiuto e tratta l'inutilità che ne risulta come un costo da assorbire in silenzio. Riconoscere la cautela eccessiva come una modalità di fallimento che ri-

chiede correzione riflette l'esperienza operativa piuttosto che una posizione difensiva reputazionale, ed è un'impostazione che gli standard specifici per settore dovrebbero mutuare.

L'appendice di valutazione, infine, rappresenta un autentico tentativo di rendere verificabili le affermazioni comportamentali. Microsoft individua quindici comportamenti, li scompone in sottocomportamenti come unità diagnostica, e analizza coppie di risposte allineate e disallineate attraverso scenari che coprono coerenza dell'identità, supervisione umana, agevolazione delle decisioni, evitamento di supposizioni, rispetto dei confini, non-inganno, falsa rassicurazione e neutralità elettorale. Il metodo di scomposizione è valido, e le coppie illustrative sono autenticamente istruttive su dove si collochi il confine. I limiti di questa appendice, discussi più avanti, sono limiti di maturità più che di approccio.

Dove la precisione operativa viene meno

The document's central weakness is that it becomes vague precisely where a procurement instrument requires specificity. Harm is the clearest example. The text acknowledges that the term is broad and context-dependent, and instructs models to weigh severity, reversibility, directness of contribution, and likelihood of deception before responding. Those are the right variables. What the document never supplies is any method for assigning values to them, any threshold at which a given combination changes the required response, or any artifact through which an institution could verify that the weighing occurred as described. A procurement officer cannot write a contract clause around proportionality that has no published scale, cannot audit adherence after deployment, and cannot distinguish a system that reasons carefully about severity from one that produces text asserting that it did.

Operator configurability compounds the problem. Microsoft states that absolute constraints cannot be overridden by operator configuration or user instruction, which is the correct commitment, but the document also situates operator behavior within applicable law, the broader governing framework, contractual terms, service-specific policies, and any other agreements operators enter with Microsoft. Those agreements are private and negotiable. The document further contemplates a set of specialized domains, including defensive cybersecurity, public safety, national security, and dual-use scientific research, where capabilities outside ordinary configurability may be authorized through separate internal review. That carve-out may well be necessary, and



Microsoft is more transparent about its existence than most vendors would be, but its practical effect is that the published constraint set is not the operative constraint set for every deployment. An institution procuring a configured deployment cannot tell from this document alone which version of the model's behavioral envelope it is buying, and the document offers no disclosure mechanism that would let it ask.

The evaluation work carries a third and more immediate limitation. Microsoft states that its models are not yet trained on the Code of Conduct, that the evaluation program is under development, and that the illustrative scenarios are synthetic, conversational rather than agentic or multimodal, and generated using one of its own models, with the misaligned examples produced by prompting that model to fail in specified ways. This is a reasonable way to bootstrap an evaluation set and an unreasonable basis for institutional assurance, because a model's own generations of good and bad behavior encode the same assumptions the model was trained on. Nothing in the appendix constitutes evidence that a deployed system behaves as specified, and Microsoft does not claim that it does. Independent adversarial testing, expert human review by domain specialists, and longitudinal measurement across real interactions are all acknowledged as future work. Until that work exists and is published, the responsible procurement position is that the behavioral claims in this document are unvalidated.

Il divario settoriale per i sistemi di lingua dei segni

Per le istituzioni che implementano intelligenza artificiale per la lingua dei segni, il problema più rilevante è che il documento è per progettazione indipendente dal settore, e i danni specifici della tecnologia della lingua dei segni non compaiono quindi da nessuna parte in esso. I vincoli assoluti riguardano armi, cyberoperazioni, perdita di controllo, manipolazione su larga scala, risposta alle crisi, impersonificazione, sicurezza dei minori, discriminazione, contenuti espliciti e violenza. Nulla in quell'elenco, e nulla nelle linee guida operative che lo seguono, arriva a toccare le modalità di fallimento che effettivamente determinano se un sistema di lingua dei segni sia sicuro da proporre a persone Deaf e DeafBlind.

Linguistic validity is the first of these. A sign language generation system can produce output that is fluent in appearance and invalid in substance, violating the grammatical structure of the target language, misusing non-manual markers that carry syntactic weight, collapsing register distinctions, or rendering regional and community variation into a synthetic standard that no community actually uses. The Code of Conduct's

accuracy provisions address factual correctness, sourcing, and calibrated confidence, none of which detect a grammatically malformed signed utterance. A system can satisfy every honesty requirement in Part 3 while producing language that a fluent signer would recognize immediately as wrong, and the framework contains no instrument that would surface the failure.

The harms that follow from that failure are equally absent. Deployment in educational settings exposes language-acquiring children to malformed input, with consequences for fluency development that no general safety category captures. Deployment in interpreted professional and civic settings can misrepresent a Deaf person′s speech to a hearing audience, with reputational and legal consequences borne entirely by the person whose language was altered. Deployment in public accommodation contexts can generate the appearance of access while delivering none, which is a compliance harm as much as a linguistic one. These are ordinary, foreseeable outcomes of the technology, and a governance document that does not name them cannot be the governance document for it.

Due ulteriori lacune sono strutturali piuttosto che linguistiche. Il documento organizza l′autorità attorno a Microsoft, agli operatori e agli utenti, e non prevede alcuna categoria per la comunità linguistica la cui lingua il sistema riproduce. Per i sistemi di lingua dei segni, l′autorità comunitaria non è una cortesia etica sovrapposta alla convalida tecnica; è la base stessa della convalida, poiché i segnanti fluenti sono l′unica fonte di verità di base su se un output sia accettabile. Un quadro normativo che riconosce solo l′istituzione acquirente e l′utente finale non ha alcun posto in cui collocare tale autorità né alcun meccanismo per richiederla. In modo correlato, il documento tratta l′accessibilità come un argomento che il modello dovrebbe gestire bene piuttosto che come una proprietà che il sistema stesso deve possedere. Se l′interfaccia, i meccanismi di supervisione, la divulgazione dello stato di IA e i percorsi di escalation siano essi stessi utilizzabili da persone Deaf e DeafBlind è una domanda che il testo non pone mai, e un sistema di lingua dei segni che fallisce su questo punto ha fallito nel punto centrale dell′esercizio.

Questioni strutturali che il documento lascia aperte

Oltre al divario settoriale, tre questioni strutturali saranno rilevanti per chiunque faccia affidamento su questo quadro normativo. La prima riguarda il pluralismo. Microsoft si impegna a sostenere un′ampia gamma di visioni del mondo e afferma che il plu-

ralismo non significa neutralità rispetto al danno, il che è la qualificazione corretta. Il quadro normativo poggia tuttavia su una posizione valoriale specifica e identificabile, incentrata sull'autonomia individuale, il consenso informato, l'agentività personale e una concezione del fiorire umano radicata nelle tradizioni dei diritti liberali, e presenta tale posizione come la base inclusiva all'interno della quale operano i valori plurali. Non c'è nulla di sbagliato nel costruire un quadro di governance su valori dichiarati. La difficoltà è che dichiarare il pluralismo mentre si operazionalizza una sola tradizione oscura la scelta alle istituzioni che la ereditano, e alcune di tali istituzioni servono comunità le cui norme di governance, in particolare riguardo all'autorità collettiva sul patrimonio culturale e linguistico condiviso, non si riducono in modo lineare alla preferenza individuale dell'utente.

La seconda riguarda la responsabilità. Si afferma che gli operatori assumono la responsabilità delle proprie configurazioni e utilizzi, ma il documento non specifica alcun obbligo di divulgazione, alcun diritto di audit, né alcun meccanismo di arbitrato per le controversie tra un operatore e le persone che la sua implementazione riguarda. Una responsabilità affermata senza una via procedurale per farla valere è una dichiarazione di intenti. Le istituzioni che fanno affidamento su questo documento dovrebbero presumere che qualunque responsabilizzazione di cui abbiano bisogno dovrà essere costruita nei propri contratti.

The third concerns conflict resolution. The framework accepts that its objectives will sometimes compete, and directs models to consult the operational guidelines, state limits to the user, propose alternatives, and, where nothing else resolves the tension, act on a holistic reading of the document as a whole. As a design philosophy for a model under uncertainty, that is defensible and probably unavoidable. As a specification, it means the resolution of genuine value conflicts is delegated to model judgment and is neither predictable in advance nor reviewable afterward. For low-stakes interactions this is acceptable. For a system mediating a Deaf person's access to medical care, employment, or legal process, it is not.

Che cosa significa questo per gli acquisti

Un'istituzione che acquista intelligenza artificiale per la lingua dei segni può utilizzare questo documento come architettura di riferimento e non dovrebbe utilizzarlo come specifica. La catena di comando, il trattamento bidirezionale del fallimento della sicurezza e la scomposizione dei comportamenti in sottocomportamenti verificabili meri-

tano tutti di essere adottati. Ciò che deve essere aggiunto è tutto ciò che il quadro generale non può fornire.

That addition begins with linguistic validation. Procurement language should specify the target language and varieties, define acceptable output at the level of grammatical structure and register, require evaluation by fluent Deaf signers rather than by proxy metrics or hearing reviewers, and set the frequency and method by which linguistic quality will be measured after deployment rather than only before it. Community authority should be written into the instrument as a standing role with defined scope, not gestured at as stakeholder engagement, and should carry the ability to reject output and halt deployment rather than merely to comment on it. Harm should be operationalized in domain terms, converting Microsoft′s proportionality variables into named failure categories with thresholds an auditor can apply. Evaluation evidence should be required before deployment and should be evidence about the configured system in the intended use case, not a vendor′s general behavioral claims. Dispute resolution should be specified contractually, including what happens when an operator configuration conflicts with community judgment about linguistic acceptability. Accessibility of the system itself, including its oversight and escalation mechanisms, should be a pass or fail condition rather than a design aspiration.

Conclusione

Microsoft ha prodotto una dichiarazione seria e insolitamente trasparente di filosofia di governance, e ne ha invitato la critica durante una finestra di consultazione definita piuttosto che presentarla come definitiva. Essa riesce come dichiarazione di principi e della struttura di autorità che tali principi richiedono. Non fornisce ancora la precisione operativa, le prove convalidate, o la tassonomia dei danni specifici per settore da cui dipende l′approvvigionamento istituzionale, e non pretende di farlo.

Il lavoro che resta da fare è quello che è sempre rimasto da fare: tradurre i principi di governance in strumenti che gli addetti agli acquisti possano scrivere nei contratti, che i revisori possano applicare dopo l′implementazione, e che le comunità interessate possano invocare quando un sistema le delude. Per l′intelligenza artificiale della lingua dei segni, tale traduzione non può essere ereditata da un quadro generale. Deve essere costruita dalle persone che sanno come appare il fallimento nella lingua.

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.

Come citare

Grizzle, H. M. (2026, September 14). Governance Aspiration and Procurement Reality: Reading Microsoft's Humanist AI Code of Conduct. Novara Consulting Group. https://doi.org/10.5281/zenodo.23003720