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/.

Governance-Anspruch und Beschaffungsrealität: Microsofts Humanist AI Code of Conduct lesen

Heather Grizzle · September 14, 2026

Microsofts Humanist AI Code of Conduct ist ein Nordstern, kein Vertrag. Wo er präzise ist, wo nicht, und was Käufer hinterfragen sollten.

Inhalt

1. What the framework gets right
2. Where operational precision fails
3. The domain gap for signed language systems
4. Structural questions the document leaves open
5. What this means for procurement
6. Fazit

Novara Consulting Group Insights

Heather M. Grizzle

Microsoft AI published its Code of Conduct for MAI models on 14 September 2026 and opened a six-week public consultation on the text. The document sets out the intended behaviors, values, and guardrails for the model family that Microsoft AI produces, organized around a design philosophy the company calls Humanist AI. Two facts about its status matter more than anything in its contents. First, Microsoft states that the document is not currently used to train its models and that a revised version, informed by consultation, is intended to guide development from 2027 onward. Second, the company describes the document as a north star rather than an account of present-day model behavior, and acknowledges a gap between what it specifies and

what its models currently do. Anyone reading this text as a compliance artifact, a warranty, or a basis for institutional assurance is reading it for something it does not claim to be.

That distinction is not a technicality. Procurement officers routinely receive vendor governance documents as evidence that a system is safe to deploy, and those documents are routinely written in a register that invites the misreading. Microsoft has been unusually candid here, and the candor is worth naming before the criticism starts. The analysis that follows treats the document as what it says it is: a draft governance philosophy under consultation, offered for exactly the kind of scrutiny this article applies.

What the framework gets right

The authority structure is the strongest part of the document. Microsoft defines a chain of command in which the Code of Conduct sits above operator policies, which in turn sit above user preferences, and it designates a category of absolute constraints that neither operators nor users can configure away. The layering is coherent, and the accompanying human control requirements, which cover interruption, shutdown, scope limits, and the integrity of action traces, are specific enough to be recognizable as engineering requirements rather than sentiment. The prohibition on models obscuring their own reasoning or communicating in forms humans cannot audit is a genuine governance commitment, because it preserves the conditions under which any downstream oversight is possible at all.

The treatment of safety failure as bidirectional is the second strength. The document names both under-caution, where a model enables harm, and over-caution, where it refuses legitimate requests or demands unnecessary confirmations, and it states that the second failure will likely occur more often even though the first is more severe. Most vendor safety frameworks optimize for refusal and treat the resulting uselessness as a cost worth absorbing silently. Naming over-caution as a failure mode that requires correction reflects operational experience rather than reputational defensiveness, and it is a framing that domain-specific standards should borrow.

The evaluation appendix, finally, represents a real attempt to make behavioral claims testable. Microsoft identifies fifteen behaviors, decomposes them into sub-behaviors as the diagnostic unit, and works through paired aligned and misaligned responses across scenarios covering identity consistency, human oversight, decision facilitation,

assumption avoidance, boundary adherence, non-deception, false reassurance, and electoral neutrality. The decomposition method is sound, and the illustrative pairs are genuinely instructive about where the line falls. The limitations of this appendix, discussed below, are limitations of maturity rather than of approach.

Where operational precision fails

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, conversa-

tional 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.

The domain gap for signed language systems

For institutions deploying sign language AI, the more consequential problem is that the document is domain-agnostic by design, and the harms specific to signed language technology therefore appear nowhere in it. The absolute constraints address weapons, cyberoperations, loss of control, manipulation at scale, crisis response, impersonation, child safety, discrimination, explicit content, and violence. Nothing in that list, and nothing in the operational guidelines that follow it, reaches the failure modes that actually govern whether a signed language system is safe to put in front of Deaf and DeafBlind people.

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.

Two further gaps are structural rather than linguistic. The document organizes authority around Microsoft, operators, and users, and has no category for the language community whose language the system reproduces. For signed language systems, community authority is not an ethical courtesy layered on top of technical validation; it is the validation baseline, because fluent signers are the only source of ground truth about whether output is acceptable. A framework that recognizes only the procuring institution and the end user has no place to put that authority and no mechanism to require it. Relatedly, the document treats accessibility as a topic the model should handle well rather than a property the system itself must have. Whether the interface, the oversight mechanisms, the disclosure of AI status, and the escalation paths are themselves usable by Deaf and DeafBlind people is a question the text never asks, and a signed language system that fails it has failed at the point of the exercise.

Structural questions the document leaves open

Beyond the domain gap, three structural issues will matter to anyone relying on this framework. The first concerns pluralism. Microsoft commits to supporting a wide range of worldviews and states that pluralism does not mean neutrality toward harm, which is the right qualification. The framework nonetheless rests on a specific and identifiable value position, centered on individual autonomy, informed consent, personal agency, and a conception of flourishing rooted in liberal rights traditions, and it presents that position as the inclusive baseline within which plural values operate. There is nothing wrong with building a governance framework on declared values. The difficulty is that declaring pluralism while operationalizing one tradition obscures the choice from the institutions that inherit it, and some of those institutions serve communities whose governance norms, particularly around collective authority over shared cultural and linguistic property, do not reduce cleanly to individual user preference.

The second concerns accountability. Operators are said to assume responsibility for their configurations and uses, but the document specifies no disclosure obligation, no audit right, and no adjudication mechanism for disputes between an operator and the people its deployment affects. Responsibility asserted without a procedural route to enforce it is a statement of intent. Institutions that rely on this document should assume that any accountability they need will have to be constructed in their own contracts.

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.

What this means for procurement

An institution procuring signed language AI can use this document as a reference architecture and should not use it as a specification. The chain of command, the bidirectional treatment of safety failure, and the decomposition of behaviors into testable sub-behaviors are all worth adopting. What has to be added is everything the general framework cannot supply.

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.

Fazit

Microsoft has produced a serious and unusually transparent statement of governance philosophy, and has invited criticism of it during a defined consultation window rather than presenting it as settled. It succeeds as a statement of principles and of the authority structure those principles require. It does not yet supply the operational precision, validated evidence, or domain-specific harm taxonomy that institutional procurement depends on, and it does not claim to.

The work that remains is the work that has always remained: translating governance principles into instruments that procurement officers can write into contracts, auditors can apply after deployment, and affected communities can invoke when a system fails them. For signed language AI, that translation cannot be inherited from a general framework. It has to be built by the people who know what failure looks like in the language.

> **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.

Zitierweise

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