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/.

Aspiration en matière de gouvernance et réalité des achats : lecture du Code de conduite de l'IA humaniste de 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.

Sommaire

1. Ce que le cadre fait bien
2. Là où la précision opérationnelle fait défaut
3. L'écart de domaine pour les systèmes de langue des signes
4. Questions structurelles laissées ouvertes par le document
5. Ce que cela signifie pour les achats
6. Conclusion

Perspectives de Novara Consulting Group

Heather M. Grizzle

Microsoft AI a publié son Code de conduite pour les modèles MAI le 14 septembre 2026 et a ouvert une consultation publique de six semaines sur le texte. Le document énonce les comportements, valeurs et garde-fous prévus pour la famille de modèles que produit Microsoft AI, organisés autour d'une philosophie de conception que l'entreprise appelle IA humaniste. Deux faits concernant son statut importent plus que tout autre élément de son contenu. Premièrement, Microsoft déclare que le document n'est pas actuellement utilisé pour entraîner ses modèles et qu'une version révisée, informée par la consultation, est destinée à guider le développement à partir de 2027. Deuxièmement, l'entreprise décrit le document comme une étoile polaire plutôt qu'un

compte rendu du comportement actuel des modèles, et reconnaît un écart entre ce qu'il spécifie et ce que font actuellement ses modèles. Quiconque lit ce texte comme un artefact de conformité, une garantie, ou une base d'assurance institutionnelle le lit pour quelque chose qu'il ne prétend pas être.

Cette distinction n'est pas une simple technicité. Les responsables des achats reçoivent régulièrement des documents de gouvernance de fournisseurs comme preuve qu'un système est sûr à déployer, et ces documents sont régulièrement rédigés dans un registre qui invite à cette lecture erronée. Microsoft a fait preuve ici d'une franchise inhabituelle, et cette franchise mérite d'être soulignée avant que la critique ne commence. L'analyse qui suit traite le document pour ce qu'il dit être : une philosophie de gouvernance en projet, soumise à consultation, offerte pour exactement le type d'examen que cet article applique.

Ce que le cadre fait bien

La structure d'autorité est la partie la plus solide du document. Microsoft définit une chaîne de commandement dans laquelle le Code de conduite se situe au-dessus des politiques des opérateurs, lesquelles se situent à leur tour au-dessus des préférences des utilisateurs, et désigne une catégorie de contraintes absolues que ni les opérateurs ni les utilisateurs ne peuvent configurer pour les contourner. La superposition est cohérente, et les exigences de contrôle humain qui l'accompagnent, lesquelles couvrent l'interruption, l'arrêt, les limites de portée et l'intégrité des traces d'action, sont suffisamment précises pour être reconnaissables comme des exigences d'ingénierie plutôt que comme de simples intentions. L'interdiction faite aux modèles d'obscurcir leur propre raisonnement ou de communiquer sous des formes que les humains ne peuvent pas vérifier constitue un véritable engagement de gouvernance, car elle préserve les conditions dans lesquelles toute supervision en aval reste possible.

Le traitement de l'échec en matière de sécurité comme bidirectionnel est la deuxième force du document. Le document nomme à la fois la sous-prudence, où un modèle facilite un préjudice, et la surprudence, où il refuse des demandes légitimes ou exige des confirmations inutiles, et affirme que la seconde défaillance se produira probablement plus souvent même si la première est plus grave. La plupart des cadres de sécurité des fournisseurs optimisent pour le refus et traitent l'inutilité qui en résulte comme un coût à absorber silencieusement. Nommer la surprudence comme un mode de défaillance nécessitant une correction reflète l'expérience opérationnelle plutôt que la

défensivité réputationnelle, et c'est un cadrage que les normes spécifiques à un domaine devraient reprendre.

L'annexe d'évaluation, enfin, représente une véritable tentative de rendre les affirmations comportementales testables. Microsoft identifie quinze comportements, les décompose en sous-comportements comme unité diagnostique, et travaille à travers des réponses appariées, alignées et non alignées, dans des scénarios couvrant la cohérence identitaire, la supervision humaine, la facilitation de décision, l'évitement des suppositions, le respect des limites, la non-tromperie, le faux réconfort et la neutralité électorale. La méthode de décomposition est solide, et les paires illustratives sont véritablement instructives quant à l'endroit où se situe la limite. Les limites de cette annexe, discutées ci-dessous, sont des limites de maturité plutôt que d'approche.

Là où la précision opérationnelle fait défaut

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.

L'écart de domaine pour les systèmes de langue des signes

Pour les institutions déployant une IA de langue des signes, le problème le plus lourd de conséquences est que le document est, par conception, agnostique quant au domaine, et les préjudices spécifiques à la technologie de la langue des signes n'apparaissent donc nulle part dans celui-ci. Les contraintes absolues traitent des armes, des cyberopérations, de la perte de contrôle, de la manipulation à grande échelle, de la réponse aux crises, de l'usurpation d'identité, de la sécurité des enfants, de la discrimination, du contenu explicite et de la violence. Rien dans cette liste, et rien dans les lignes directrices opérationnelles qui la suivent, n'atteint les modes de défaillance qui déterminent réellement si un système de langue des signes est sûr à mettre devant des personnes Deaf et sourdaveugles.

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.

Deux autres lacunes sont structurelles plutôt que linguistiques. Le document organise l'autorité autour de Microsoft, des opérateurs et des utilisateurs, et ne prévoit aucune catégorie pour la communauté linguistique dont le système reproduit la langue. Pour les systèmes de langue des signes, l'autorité communautaire n'est pas une courtoisie éthique superposée à la validation technique ; c'est la référence de validation elle-même, car les signeurs fluides sont la seule source de vérité de terrain quant à l'acceptabilité de la sortie. Un cadre qui ne reconnaît que l'institution acquéreuse et l'utilisateur final n'a nulle part où placer cette autorité et aucun mécanisme pour l'exiger. De manière liée, le document traite l'accessibilité comme un sujet que le modèle devrait bien gérer plutôt que comme une propriété que le système lui-même doit posséder. Que l'interface, les mécanismes de supervision, la divulgation du statut d'IA et les voies d'escalade soient eux-mêmes utilisables par les personnes Deaf et sourdaveugles est une question que le texte ne pose jamais, et un système de langue des signes qui y échoue a échoué au point même de l'exercice.

Questions structurelles laissées ouvertes par le document

Au-delà de l'écart de domaine, trois questions structurelles importeront à quiconque s'appuie sur ce cadre. La première concerne le pluralisme. Microsoft s'engage à soutenir un large éventail de visions du monde et affirme que le pluralisme ne signifie pas

la neutralité face au préjudice, ce qui constitue la bonne nuance. Le cadre repose néanmoins sur une position de valeurs spécifique et identifiable, centrée sur l'autonomie individuelle, le consentement éclairé, l'agentivité personnelle, et une conception de l'épanouissement enracinée dans les traditions des droits libéraux, et il présente cette position comme la base inclusive au sein de laquelle opèrent les valeurs plurielles. Il n'y a rien de mal à construire un cadre de gouvernance sur des valeurs déclarées. La difficulté est que déclarer le pluralisme tout en opérationnalisant une seule tradition obscurcit ce choix aux yeux des institutions qui en héritent, et certaines de ces institutions servent des communautés dont les normes de gouvernance, en particulier autour de l'autorité collective sur un patrimoine culturel et linguistique partagé, ne se réduisent pas proprement à une préférence individuelle d'utilisateur.

La deuxième concerne la responsabilité. Il est dit que les opérateurs assument la responsabilité de leurs configurations et de leurs usages, mais le document ne précise aucune obligation de divulgation, aucun droit d'audit, et aucun mécanisme d'arbitrage pour les litiges entre un opérateur et les personnes affectées par son déploiement. Une responsabilité affirmée sans voie procédurale pour l'appliquer n'est qu'une déclaration d'intention. Les institutions qui s'appuient sur ce document devraient supposer que toute responsabilité dont elles ont besoin devra être construite dans leurs propres contrats.

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.

Ce que cela signifie pour les achats

Une institution achetant une IA de langue des signes peut utiliser ce document comme architecture de référence et ne devrait pas l'utiliser comme spécification. La chaîne de commandement, le traitement bidirectionnel de l'échec en matière de sécurité, et la décomposition des comportements en sous-comportements testables mé-

ritent tous d'être adoptés. Ce qui doit être ajouté, c'est tout ce que le cadre général ne peut pas fournir.

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.

Conclusion

Microsoft a produit un énoncé sérieux et inhabituellement transparent de philosophie de gouvernance, et a invité la critique de celui-ci pendant une fenêtre de consultation définie plutôt que de le présenter comme définitif. Il réussit en tant qu'énoncé de principes et de la structure d'autorité que ces principes exigent. Il ne fournit pas encore la précision opérationnelle, les preuves validées, ou la taxonomie des préjudices spécifiques au domaine dont dépendent les achats institutionnels, et il ne prétend pas le faire.

Le travail qui reste est le travail qui a toujours été à faire : traduire les principes de gouvernance en instruments que les responsables des achats peuvent inscrire dans des contrats, que les auditeurs peuvent appliquer après le déploiement, et que les communautés affectées peuvent invoquer lorsqu'un système leur fait défaut. Pour l'IA de langue des signes, cette traduction ne peut pas être héritée d'un cadre général. Elle doit être construite par les personnes qui savent à quoi ressemble l'échec dans la langue.

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.

Comment citer

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