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/.

Aspiracje w zakresie zarządzania a rzeczywistość zamówień publicznych: lektura Kodeksu Postępowania Humanistycznej AI firmy Microsoft

Heather Grizzle · 14 września 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.

Spis treści

1. Co w tej strukturze jest trafne
2. Gdzie zawodzi precyzja operacyjna
3. Luka dziedzinowa dotycząca systemów języka migowego
4. Kwestie strukturalne pozostawione bez odpowiedzi
5. Co to oznacza dla zamówień publicznych
6. Podsumowanie

Analizy Novara Consulting Group

Heather M. Grizzle

Microsoft AI opublikował swój Kodeks Postępowania dla modeli MAI 14 września 2026 roku i rozpoczął sześciotygodniowe konsultacje publiczne dotyczące tego tekstu. Dokument określa zamierzone zachowania, wartości i mechanizmy zabezpieczające dla rodziny modeli produkowanych przez Microsoft AI, zorganizowane wokół filozofii projektowej, którą firma nazywa Humanist AI. Dwa fakty dotyczące jego statusu są ważniejsze niż cokolwiek w jego treści. Po pierwsze, Microsoft stwierdza, że dokument nie jest obecnie wykorzystywany do trenowania jego modeli, a poprawiona wersja,

uwzględniająca wyniki konsultacji, ma kierować rozwojem od 2027 roku. Po drugie, firma opisuje dokument jako gwiazdę przewodnią, a nie opis obecnego zachowania modeli, i przyznaje, że istnieje rozbieżność między tym, co dokument określa, a tym, co jej modele robią obecnie. Każdy, kto czyta ten tekst jako artefakt zgodności, gwarancję lub podstawę instytucjonalnego zapewnienia, czyta go jako coś, czym on nie twierdzi, że jest.

To rozróżnienie nie jest kwestią techniczną. Urzędnicy ds. zamówień rutynowo otrzymują dokumenty dostawców dotyczące zarządzania jako dowód, że system jest bezpieczny do wdrożenia, a dokumenty te są rutynowo pisane w rejestrze, który zachęca do takiej błędnej interpretacji. Microsoft był w tym przypadku niezwykle szczery, i tę szczerość warto nazwać, zanim zacznie się krytyka. Poniższa analiza traktuje dokument jako to, czym sam się deklaruje: roboczą filozofię zarządzania poddaną konsultacjom, oferowaną dokładnie pod taki rodzaj badania, jaki stosuje ten artykuł.

Co w tej strukturze jest trafne

Struktura władzy jest najmocniejszą częścią dokumentu. Microsoft definiuje łańcuch podporządkowania, w którym Kodeks Postępowania znajduje się ponad politykami operatorów, te zaś ponad preferencjami użytkowników, oraz wyznacza kategorię absolutnych ograniczeń, których nie mogą zmienić ani operatorzy, ani użytkownicy. Ta warstwowość jest spójna, a towarzyszące jej wymagania dotyczące kontroli człowieka – obejmujące przerywanie działania, wyłączanie, ograniczenia zakresu i integralność śladów działań – są na tyle konkretne, że można je rozpoznać jako wymagania inżynierskie, a nie jako deklaracje intencji. Zakaz ukrywania przez modele własnego rozumowania lub komunikowania się w formach niemożliwych do zweryfikowania przez ludzi stanowi rzeczywiste zobowiązanie w zakresie zarządzania, ponieważ zachowuje warunki, w jakich jakikolwiek dalszy nadzór jest w ogóle możliwy.

Drugą mocną stroną jest traktowanie niepowodzenia w zakresie bezpieczeństwa jako zjawiska dwukierunkowego. Dokument nazywa zarówno niedostateczną ostrożność, w której model umożliwia wyrządzenie szkody, jak i nadmierną ostrożność, w której odmawia on realizacji uzasadnionych próśb lub żąda zbędnych potwierdzeń, i stwierdza, że ten drugi rodzaj niepowodzenia będzie prawdopodobnie występował częściej, mimo że pierwszy jest poważniejszy. Większość ram bezpieczeństwa stosowanych przez dostawców optymalizuje się pod kątem odmowy i traktuje wynikającą z tego bezużyteczność jako koszt, który należy po cichu zaakceptować. Nazwanie nadmiernej ostrożno-

ści błędem wymagającym korekty odzwierciedla doświadczenie operacyjne, a nie postawę obronną nastawioną na reputację, i jest to ujęcie, które powinny zapożyczyć standardy specyficzne dla poszczególnych dziedzin.

Wreszcie załącznik dotyczący ewaluacji stanowi rzeczywistą próbę uczynienia twierdzeń dotyczących zachowania możliwymi do przetestowania. Microsoft identyfikuje piętnaście zachowań, rozkłada je na podzachowania jako jednostkę diagnostyczną i analizuje pary odpowiedzi zgodnych i niezgodnych z wytycznymi w scenariuszach obejmujących spójność tożsamości, nadzór człowieka, ułatwianie podejmowania decyzji, unikanie założeń, przestrzeganie granic, niestosowanie oszustwa, fałszywe uspokajanie oraz neutralność wyborczą. Metoda dekompozycji jest solidna, a przykładowe pary są naprawdę pouczające co do tego, gdzie przebiega granica. Ograniczenia tego załącznika, omówione poniżej, wynikają raczej z niedojrzałości niż z wadliwego podejścia.

Gdzie zawodzi precyzja operacyjna

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 au-

thorized 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.

Luka dziedzinowa dotycząca systemów języka migowego

Dla instytucji wdrażających sztuczną inteligencję języka migowego poważniejszym problemem jest to, że dokument jest z założenia neutralny dziedzinowo, a zagrożenia specyficzne dla technologii języka migowego nie pojawiają się w nim nigdzie. Absolutne ograniczenia dotyczą broni, cyberoperacji, utraty kontroli, manipulacji na dużą skalę, reagowania kryzysowego, podszywania się, bezpieczeństwa dzieci, dyskryminacji, treści jednoznacznie seksualnych oraz przemocy. Nic na tej liście, ani nic w następujących po niej wytycznych operacyjnych, nie sięga do trybów awarii, które faktycznie decydują o tym, czy system języka migowego jest bezpieczny, by postawić go przed osobami Głuchymi i głuchoniewidomymi.

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 varia-

tion 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.

Dwie dalsze luki mają charakter strukturalny, a nie językowy. Dokument organizuje strukturę władzy wokół Microsoft, operatorów i użytkowników, nie przewidując żadnej kategorii dla społeczności językowej, której język system odtwarza. W przypadku systemów języka migowego autorytet społeczności nie jest grzecznościowym elementem etycznym nałożonym na walidację techniczną; jest on podstawą samej walidacji, ponieważ biegli użytkownicy języka migowego są jedynym źródłem prawdy odniesienia co do tego, czy wynik jest akceptowalny. Struktura uznająca jedynie instytucję zamawiającą i użytkownika końcowego nie ma miejsca, by umieścić ten autorytet, ani mechanizmu, by go wymagać. Podobnie dokument traktuje dostępność jako temat, z którym model powinien sobie dobrze radzić, a nie jako właściwość, którą musi posiadać sam system. Czy interfejs, mechanizmy nadzoru, ujawnianie statusu sztucznej inteligencji oraz ścieżki eskalacji są same w sobie użyteczne dla osób Głuchych i głuchoniewidomych – to pytanie, którego tekst nigdy nie zadaje, a system języka migowego, który go nie spełnia, zawiódł w samym sednie tego przedsięwzięcia.

Kwestie strukturalne pozostawione bez odpowiedzi

Poza luką dziedzinową, trzy kwestie strukturalne będą miały znaczenie dla każdego, kto polega na tej strukturze. Pierwsza dotyczy pluralizmu. Microsoft zobowiązuje się

do wspierania szerokiego zakresu światopoglądów i stwierdza, że pluralizm nie oznacza neutralności wobec szkody, co jest właściwym zastrzeżeniem. Struktura ta opiera się jednak na konkretnym i możliwym do zidentyfikowania stanowisku wartości, skoncentrowanym na autonomii jednostki, świadomej zgodzie, osobistej sprawczości oraz koncepcji dobrostanu zakorzenionej w liberalnych tradycjach praw, i przedstawia to stanowisko jako inkluzywną podstawę, w ramach której funkcjonują wartości pluralne. Nie ma nic złego w budowaniu struktury zarządzania na deklarowanych wartościach. Trudność polega na tym, że deklarowanie pluralizmu przy jednoczesnym operacjonalizowaniu jednej tradycji zaciemnia ten wybór przed instytucjami, które go dziedziczą, a niektóre z tych instytucji służą społecznościom, których normy zarządzania, szczególnie w odniesieniu do zbiorowego autorytetu nad wspólnym dziedzictwem kulturowym i językowym, nie dają się sprowadzić wprost do indywidualnych preferencji użytkownika.

Druga kwestia dotyczy odpowiedzialności. Mówi się, że operatorzy biorą na siebie odpowiedzialność za swoje konfiguracje i zastosowania, ale dokument nie określa żadnego obowiązku ujawniania, żadnego prawa do kontroli ani żadnego mechanizmu rozstrzygania sporów między operatorem a osobami, na które wpływa jego wdrożenie. Odpowiedzialność deklarowana bez proceduralnej ścieżki do jej egzekwowania jest jedynie wyrazem intencji. Instytucje polegające na tym dokumencie powinny założyć, że wszelka potrzebna im odpowiedzialność będzie musiała zostać skonstruowana we własnych umowach.

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.

Co to oznacza dla zamówień publicznych

Instytucja nabywająca sztuczną inteligencję języka migowego może wykorzystać ten dokument jako architekturę odniesienia i nie powinna wykorzystywać go jako specyfi-

kacji. Łańcuch podporządkowania, dwukierunkowe traktowanie niepowodzenia w zakresie bezpieczeństwa oraz dekompozycja zachowań na możliwe do przetestowania podzachowania – wszystko to warto przyjąć. Trzeba dodać wszystko to, czego ogólna struktura nie jest w stanie dostarczyć.

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.

Podsumowanie

Microsoft przygotował poważne i niezwykle przejrzyste oświadczenie dotyczące filozofii zarządzania i zaprosił do jego krytyki w ramach określonego okresu konsultacji, zamiast przedstawiać je jako sprawę zamkniętą. Udaje mu się to jako deklaracja zasad i struktury władzy, jakiej te zasady wymagają. Nie dostarcza jeszcze precyzji operacyjnej, zwalidowanych dowodów ani dziedzinowej taksonomii szkód, od których zależą instytucjonalne zamówienia publiczne, i nie twierdzi, że to robi.

Pozostająca do wykonania praca to praca, która zawsze pozostawała do wykonania: przełożenie zasad zarządzania na instrumenty, które urzędnicy ds. zamówień mogą zapisać w umowach, audytorzy mogą zastosować po wdrożeniu, a dotknięte społeczności mogą przywołać, gdy system je zawiedzie. W przypadku sztucznej inteligencji języka migowego tego przełożenia nie można odziedziczyć z ogólnej struktury. Musi je zbudować ci, którzy wiedzą, jak wygląda niepowodzenie w danym języku.

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.

Jak cytować

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