Este documento foi traduzido do inglês por máquina e pode conter erros. O original em inglês está em https://www.novaracg.com/2026/09/14/microsofts-humanist-ai-code-of-conduct-governance-aspiration-and-procurement-reality/.

Aspiração de Governança e Realidade de Aquisição: Lendo o Código de Conduta de IA Humanista da Microsoft

Heather Grizzle · 14 de setembro de 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.

Conteúdo

1. O que o framework acerta
2. Onde a precisão operacional falha
3. A lacuna de domínio para sistemas de língua de sinais
4. Questões estruturais que o documento deixa em aberto
5. O que isso significa para aquisições
6. Conclusão

Insights da Novara Consulting Group

Heather M. Grizzle

A Microsoft AI publicou seu Código de Conduta para modelos MAI em 14 de setembro de 2026 e abriu uma consulta pública de seis semanas sobre o texto. O documento estabelece os comportamentos, valores e salvaguardas pretendidos para a família de modelos que a Microsoft AI produz, organizados em torno de uma filosofia de design que a empresa chama de IA Humanista. Dois fatos sobre seu status importam mais do que qualquer coisa em seu conteúdo. Primeiro, a Microsoft afirma que o documento não é atualmente usado para treinar seus modelos e que uma versão revisada, informada pela consulta, pretende orientar o desenvolvimento a partir de 2027. Segundo, a empresa descreve o documento como uma estrela-guia em vez de um retrato do com-

portamento atual dos modelos, e reconhece uma lacuna entre o que ele especifica e o que seus modelos realmente fazem hoje. Quem lê este texto como um artefato de conformidade, uma garantia ou uma base para segurança institucional está lendo-o em busca de algo que ele não pretende ser.

Essa distinção não é uma questão técnica menor. Os responsáveis por aquisições costumam receber documentos de governança de fornecedores como prova de que um sistema é seguro para implantação, e esses documentos costumam ser escritos em um registro que convida a essa leitura equivocada. A Microsoft foi incomumente franca aqui, e essa franqueza merece ser reconhecida antes que a crítica comece. A análise a seguir trata o documento como aquilo que ele afirma ser: uma filosofia de governança em rascunho, em consulta, oferecida exatamente para o tipo de escrutínio que este artigo aplica.

O que o framework acerta

A estrutura de autoridade é a parte mais forte do documento. A Microsoft define uma cadeia de comando na qual o Código de Conduta se situa acima das políticas do operador, que por sua vez se situam acima das preferências do usuário, e designa uma categoria de restrições absolutas que nem operadores nem usuários podem configurar para contornar. A estratificação é coerente, e os requisitos de controle humano que a acompanham, que abrangem interrupção, desligamento, limites de escopo e a integridade dos registros de ação, são específicos o suficiente para serem reconhecíveis como requisitos de engenharia, e não como boas intenções. A proibição de que os modelos ocultem seu próprio raciocínio ou se comuniquem de formas que os humanos não possam auditar é um compromisso de governança genuíno, pois preserva as condições sob as quais qualquer supervisão posterior é sequer possível.

O tratamento da falha de segurança como bidirecional é o segundo ponto forte. O documento nomeia tanto a cautela insuficiente, em que um modelo possibilita danos, quanto a cautela excessiva, em que ele recusa solicitações legítimas ou exige confirmações desnecessárias, e afirma que a segunda falha provavelmente ocorrerá com mais frequência, ainda que a primeira seja mais grave. A maioria dos frameworks de segurança de fornecedores otimiza para a recusa e trata a inutilidade resultante como um custo a ser absorvido silenciosamente. Nomear a cautela excessiva como um modo de falha que exige correção reflete experiência operacional em vez de defensividade reputacional, e é um enquadramento que padrões específicos de domínio deveriam adotar.

O apêndice de avaliação, por fim, representa uma tentativa real de tornar testáveis as afirmações comportamentais. A Microsoft identifica quinze comportamentos, decompõe-nos em subcomportamentos como unidade de diagnóstico, e trabalha com pares de respostas alinhadas e desalinhadas em cenários que cobrem consistência de identidade, supervisão humana, facilitação de decisões, evitação de pressupostos, adesão a limites, não-decepção, falsa tranquilização e neutralidade eleitoral. O método de decomposição é sólido, e os pares ilustrativos são genuinamente instrutivos sobre onde a linha se situa. As limitações deste apêndice, discutidas abaixo, são limitações de maturidade, não de abordagem.

Onde a precisão operacional falha

A fraqueza central do documento é que ele se torna vago precisamente onde um instrumento de aquisição exige especificidade. O dano é o exemplo mais claro. O texto reconhece que o termo é amplo e dependente do contexto, e instrui os modelos a ponderar a gravidade, a reversibilidade, a diretividade da contribuição e a probabilidade de engano antes de responder. Essas são as variáveis certas. O que o documento nunca fornece é qualquer método para atribuir valores a elas, qualquer limiar no qual uma determinada combinação altere a resposta exigida, ou qualquer artefato através do qual uma instituição pudesse verificar que a ponderação ocorreu conforme descrito. Um responsável por aquisições não pode redigir uma cláusula contratual em torno de uma proporcionalidade que não tem escala publicada, não pode auditar a conformidade após a implementação e não pode distinguir um sistema que raciocina cuidadosamente sobre a gravidade de um que produz texto afirmando que o fez.

A configurabilidade pelo operador agrava o problema. A Microsoft afirma que as restrições absolutas não podem ser substituídas pela configuração do operador ou pela instrução do usuário, o que é o compromisso correto, mas o documento também situa o comportamento do operador dentro da lei aplicável, do quadro de governança mais amplo, dos termos contratuais, das políticas específicas do serviço e de quaisquer outros acordos que os operadores celebrem com a Microsoft. Esses acordos são privados e negociáveis. O documento ainda contempla um conjunto de domínios especializados, incluindo cibersegurança defensiva, segurança pública, segurança nacional e pesquisa científica de duplo uso, onde capacidades fora da configurabilidade comum podem ser autorizadas através de revisão interna separada. Essa exceção pode muito bem ser necessária, e a Microsoft é mais transparente sobre a sua existência do que a

maioria dos fornecedores seria, mas o seu efeito prático é que o conjunto de restrições publicado não é o conjunto de restrições operativo para todas as implementações. Uma instituição que adquire uma implementação configurada não consegue determinar apenas a partir deste documento qual versão do envelope comportamental do modelo está adquirindo, e o documento não oferece nenhum mecanismo de divulgação que lhe permitisse perguntar.

O trabalho de avaliação carrega uma terceira limitação, e mais imediata. A Microsoft afirma que os seus modelos ainda não foram treinados no Código de Conduta, que o programa de avaliação está em desenvolvimento e que os cenários ilustrativos são sintéticos, conversacionais em vez de agênticos ou multimodais, e gerados usando um dos seus próprios modelos, com os exemplos desalinhados produzidos ao instigar esse modelo a falhar de formas especificadas. Esta é uma forma razoável de dar início a um conjunto de avaliação e uma base pouco razoável para garantia institucional, porque as próprias gerações de um modelo de comportamento bom e mau codificam os mesmos pressupostos com que o modelo foi treinado. Nada no apêndice constitui prova de que um sistema implementado se comporta conforme especificado, e a Microsoft não afirma que sim. Testes adversariais independentes, revisão humana especializada por peritos de domínio e medição longitudinal ao longo de interações reais são todos reconhecidos como trabalho futuro. Até que esse trabalho exista e seja publicado, a posição responsável em matéria de aquisição é a de que as afirmações comportamentais deste documento não estão validadas.

A lacuna de domínio para sistemas de língua de sinais

Para instituições que implantam IA de língua de sinais, o problema mais consequente é que o documento é, por design, agnóstico de domínio, e os danos específicos à tecnologia de língua de sinais, portanto, não aparecem em nenhum lugar nele. As restrições absolutas abordam armas, ciberoperações, perda de controle, manipulação em escala, resposta a crises, personificação, segurança infantil, discriminação, conteúdo explícito e violência. Nada nessa lista, e nada nas diretrizes operacionais que a seguem, alcança os modos de falha que realmente determinam se um sistema de língua de sinais é seguro para ser colocado diante de pessoas Deaf e DeafBlind.

A validade linguística é a primeira destas questões. Um sistema de geração de língua gestual pode produzir resultados fluentes na aparência e inválidos na substância, violando a estrutura gramatical da língua-alvo, utilizando incorretamente marcadores

não manuais que carregam peso sintático, colapsando distinções de registo, ou transformando a variação regional e comunitária num padrão sintético que nenhuma comunidade utiliza de facto. As disposições de exatidão do Código de Conduta abordam a correção factual, a citação de fontes e a confiança calibrada, nenhuma das quais deteta um enunciado gestual gramaticalmente malformado. Um sistema pode satisfazer todos os requisitos de honestidade da Parte 3 ao mesmo tempo que produz uma língua que um sinalizante fluente reconheceria imediatamente como errada, e o quadro de governança não contém nenhum instrumento que faria essa falha emergir.

Os danos que decorrem dessa falha estão igualmente ausentes. A implementação em contextos educativos expõe crianças em fase de aquisição da língua a input malformado, com consequências para o desenvolvimento da fluência que nenhuma categoria geral de segurança capta. A implementação em contextos profissionais e cívicos interpretados pode deturpar o discurso de uma pessoa Surda perante um público ouvinte, com consequências reputacionais e legais suportadas inteiramente pela pessoa cuja língua foi alterada. A implementação em contextos de acomodação pública pode gerar a aparência de acesso sem oferecer nenhum, o que constitui um dano de conformidade tanto quanto um dano linguístico. Estes são resultados comuns e previsíveis da tecnologia, e um documento de governança que não os nomeia não pode ser o documento de governança para ela.

Duas outras lacunas são estruturais em vez de linguísticas. O documento organiza a autoridade em torno da Microsoft, dos operadores e dos usuários, e não tem nenhuma categoria para a comunidade linguística cuja língua o sistema reproduz. Para sistemas de língua de sinais, a autoridade comunitária não é uma cortesia ética sobreposta à validação técnica; é a linha de base de validação, porque sinalizantes fluentes são a única fonte de verdade fundamental sobre se a saída é aceitável. Um framework que reconhece apenas a instituição adquirente e o usuário final não tem onde colocar essa autoridade nem mecanismo para exigi-la. Relacionadamente, o documento trata a acessibilidade como um tópico que o modelo deve tratar bem, em vez de uma propriedade que o próprio sistema deve possuir. Se a interface, os mecanismos de supervisão, a divulgação do status de IA e os caminhos de escalonamento são eles próprios utilizáveis por pessoas Deaf e DeafBlind é uma pergunta que o texto nunca faz, e um sistema de língua de sinais que falha nisso falhou no ponto central do exercício.

Questões estruturais que o documento deixa em aberto

Além da lacuna de domínio, três questões estruturais importarão a quem confiar neste framework. A primeira diz respeito ao pluralismo. A Microsoft se compromete a apoiar uma ampla gama de visões de mundo e afirma que pluralismo não significa neutralidade diante do dano, o que é a qualificação correta. Ainda assim, o framework repousa sobre uma posição de valores específica e identificável, centrada na autonomia individual, no consentimento informado, na capacidade de agência pessoal e em uma concepção de florescimento enraizada nas tradições liberais de direitos, e apresenta essa posição como a linha de base inclusiva dentro da qual valores plurais operam. Não há nada de errado em construir um framework de governança sobre valores declarados. A dificuldade é que declarar pluralismo ao operacionalizar uma única tradição obscurece a escolha para as instituições que a herdam, e algumas dessas instituições servem comunidades cujas normas de governança, particularmente em torno da autoridade coletiva sobre propriedade cultural e linguística compartilhada, não se reduzem facilmente à preferência individual do usuário.

A segunda diz respeito à responsabilização. Diz-se que os operadores assumem responsabilidade por suas configurações e usos, mas o documento não especifica nenhuma obrigação de divulgação, nenhum direito de auditoria, e nenhum mecanismo de arbitragem para disputas entre um operador e as pessoas afetadas por sua implantação. Responsabilidade afirmada sem uma via processual para fazê-la valer é uma declaração de intenção. As instituições que confiam neste documento devem presumir que qualquer responsabilização de que precisem terá de ser construída em seus próprios contratos.

O terceiro diz respeito à resolução de conflitos. O quadro aceita que os seus objetivos irão por vezes competir entre si, e instrui os modelos a consultar as diretrizes operacionais, declarar os limites ao utilizador, propor alternativas e, quando mais nada resolve a tensão, agir com base numa leitura holística do documento como um todo. Como filosofia de conceção para um modelo em situação de incerteza, isso é defensável e provavelmente inevitável. Como especificação, significa que a resolução de conflitos de valores genuínos é delegada ao julgamento do modelo e não é previsível de antemão nem revisável posteriormente. Para interações de baixo risco, isto é aceitável. Para um sistema que medeia o acesso de uma pessoa Surda a cuidados médicos, emprego ou processo jurídico, não é.

O que isso significa para aquisições

Uma instituição que adquire IA de língua de sinais pode usar este documento como uma arquitetura de referência e não deveria usá-lo como uma especificação. A cadeia de comando, o tratamento bidirecional da falha de segurança e a decomposição de comportamentos em subcomportamentos testáveis merecem todos ser adotados. O que precisa ser acrescentado é tudo aquilo que o framework geral não consegue fornecer.

Esse acréscimo começa com a validação linguística. A linguagem de aquisição deve especificar a língua-alvo e as suas variedades, definir o resultado aceitável ao nível da estrutura gramatical e do registo, exigir avaliação por sinalizantes Surdos fluentes em vez de métricas por procuração ou revisores ouvintes, e estabelecer a frequência e o método pelos quais a qualidade linguística será medida após a implementação, e não apenas antes dela. A autoridade comunitária deve ser inscrita no instrumento como um papel permanente com âmbito definido, e não meramente aludida como envolvimento das partes interessadas, e deve ter a capacidade de rejeitar resultados e suspender a implementação, em vez de apenas comentar sobre eles. O dano deve ser operacionalizado em termos de domínio, convertendo as variáveis de proporcionalidade da Microsoft em categorias de falha nomeadas com limiares que um auditor possa aplicar. As provas de avaliação devem ser exigidas antes da implementação e devem constituir prova sobre o sistema configurado no caso de uso pretendido, e não afirmações comportamentais gerais de um fornecedor. A resolução de disputas deve ser especificada contratualmente, incluindo o que acontece quando uma configuração do operador entra em conflito com o julgamento da comunidade sobre a aceitabilidade linguística. A acessibilidade do próprio sistema, incluindo os seus mecanismos de supervisão e escalonamento, deve ser uma condição de aprovação ou reprovação, e não uma aspiração de conceção.

Conclusão

A Microsoft produziu uma declaração séria e incomumente transparente de filosofia de governança, e convidou críticas a ela durante uma janela de consulta definida, em vez de apresentá-la como algo já decidido. Ela é bem-sucedida como declaração de princípios e da estrutura de autoridade que esses princípios exigem. Ela ainda não fornece a precisão operacional, a evidência validada ou a taxonomia de danos específica de domínio de que a aquisição institucional depende, e não afirma fazê-lo.

O trabalho que resta é o trabalho que sempre restou: traduzir princípios de governança em instrumentos que os responsáveis por aquisições possam redigir em contratos, que auditores possam aplicar após a implantação, e que comunidades afetadas possam invocar quando um sistema as falhar. Para a IA de língua de sinais, essa tradução não pode ser herdada de um framework geral. Ela precisa ser construída pelas pessoas que sabem como é a falha naquela língua.

> **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.

Como citar

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