NOVARA
CONSULTING GROUP

Este documento fue traducido del inglés por una máquina y puede contener errores. El original en inglés está disponible en https://www.novaracg.com/2026/09/14/microsofts-humanist-ai-code-of-conduct-governance-aspiration-and-procurement-reality/.

Aspiración de gobernanza y realidad de las adquisiciones: una lectura del Código de Conducta de IA Humanista de Microsoft

Heather Grizzle · 14 de septiembre 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.

Contenidos

1. Lo que el marco hace bien
2. Dónde falla la precisión operativa
3. La brecha de dominio para los sistemas de lengua de señas
4. Cuestiones estructurales que el documento deja abiertas
5. Lo que esto significa para las adquisiciones
6. Conclusión

Perspectivas de Novara Consulting Group

Heather M. Grizzle

Microsoft AI publicó su Código de Conducta para los modelos MAI el 14 de septiembre de 2026 y abrió una consulta pública de seis semanas sobre el texto. El documento establece los comportamientos, valores y salvaguardas previstos para la familia de modelos que produce Microsoft AI, organizados en torno a una filosofía de diseño que la empresa denomina Humanist AI. Dos hechos sobre su estatus importan más que cualquier cosa en su contenido. Primero, Microsoft afirma que el documento no se utiliza actualmente para entrenar sus modelos y que una versión revisada, informada por

la consulta, está destinada a guiar el desarrollo a partir de 2027. Segundo, la empresa describe el documento como una estrella polar más que como un relato del comportamiento actual de los modelos, y reconoce una brecha entre lo que este especifica y lo que sus modelos hacen en la actualidad. Cualquiera que lea este texto como un artefacto de cumplimiento, una garantía o una base para la certeza institucional lo está leyendo como algo que no pretende ser.

Esa distinción no es una cuestión técnica menor. Los responsables de adquisiciones reciben habitualmente documentos de gobernanza de proveedores como prueba de que un sistema es seguro para su implementación, y esos documentos suelen redactarse en un registro que invita a esa interpretación errónea. Microsoft ha sido inusualmente franca en este caso, y esa franqueza merece reconocerse antes de que comiencen las críticas. El análisis que sigue trata el documento como lo que dice ser: una filosofía de gobernanza en borrador sometida a consulta, ofrecida precisamente para el tipo de escrutinio que este artículo aplica.

Lo que el marco hace bien

La estructura de autoridad es la parte más sólida del documento. Microsoft define una cadena de mando en la que el Código de Conducta se sitúa por encima de las políticas de los operadores, que a su vez se sitúan por encima de las preferencias de los usuarios, y designa una categoría de restricciones absolutas que ni los operadores ni los usuarios pueden desactivar mediante configuración. La estructura por capas es coherente, y los requisitos de control humano que la acompañan, que abarcan la interrupción, el apagado, los límites de alcance y la integridad de los registros de acciones, son suficientemente específicos como para reconocerse como requisitos de ingeniería y no como declaraciones de intención. La prohibición de que los modelos oculten su propio razonamiento o se comuniquen en formas que los humanos no puedan auditar es un compromiso de gobernanza genuino, porque preserva las condiciones bajo las cuales cualquier supervisión posterior es siquiera posible.

El tratamiento de la falla de seguridad como bidireccional es la segunda fortaleza. El documento nombra tanto la subcautela, en la que un modelo facilita un daño, como la sobrecautela, en la que rechaza solicitudes legítimas o exige confirmaciones innecesarias, y afirma que la segunda falla probablemente ocurrirá con más frecuencia aunque la primera sea más grave. La mayoría de los marcos de seguridad de los proveedores optimizan para el rechazo y tratan la inutilidad resultante como un costo que se absor-

be en silencio. Nombrar la sobrecautela como un modo de falla que requiere corrección refleja experiencia operativa más que una postura defensiva de reputación, y es un enfoque que las normas específicas de dominio deberían adoptar.

El apéndice de evaluación, por último, representa un intento real de hacer verificables las afirmaciones sobre el comportamiento. Microsoft identifica quince comportamientos, los descompone en subcomportamientos como unidad de diagnóstico, y desarrolla pares de respuestas alineadas y desalineadas en escenarios que cubren la consistencia de identidad, la supervisión humana, la facilitación de decisiones, la evitación de supuestos, el respeto de límites, la no engaño, la falsa tranquilización y la neutralidad electoral. El método de descomposición es sólido, y los pares ilustrativos son genuinamente instructivos sobre dónde se traza la línea. Las limitaciones de este apéndice, discutidas más abajo, son limitaciones de madurez y no de enfoque.

Dónde falla la precisión operativa

La debilidad central del documento es que se vuelve impreciso precisamente donde un instrumento de contratación requiere especificidad. El daño es el ejemplo más claro. El texto reconoce que el término es amplio y depende del contexto, e instruye a los modelos a sopesar la gravedad, la reversibilidad, la directitud de la contribución y la probabilidad de engaño antes de responder. Esas son las variables correctas. Lo que el documento nunca proporciona es método alguno para asignarles valores, umbral alguno en el que una combinación dada cambie la respuesta requerida, ni artefacto alguno mediante el cual una institución pueda verificar que la ponderación se produjo tal como se describe. Un responsable de contratación no puede redactar una cláusula contractual sobre proporcionalidad que no tenga una escala publicada, no puede auditar el cumplimiento después de la implementación, y no puede distinguir un sistema que razona cuidadosamente sobre la gravedad de uno que produce texto que afirma haberlo hecho.

La configurabilidad del operador agrava el problema. Microsoft afirma que las restricciones absolutas no pueden ser anuladas por la configuración del operador ni por instrucciones del usuario, lo cual es el compromiso correcto, pero el documento también sitúa el comportamiento del operador dentro de la ley aplicable, el marco de gobernanza más amplio, los términos contractuales, las políticas específicas del servicio y cualquier otro acuerdo que los operadores celebren con Microsoft. Esos acuerdos son privados y negociables. El documento además contempla un conjunto de dominios es-

pecializados, incluida la ciberseguridad defensiva, la seguridad pública, la seguridad nacional y la investigación científica de doble uso, donde capacidades fuera de la configurabilidad ordinaria pueden autorizarse mediante una revisión interna separada. Esa excepción bien puede ser necesaria, y Microsoft es más transparente sobre su existencia de lo que serían la mayoría de los proveedores, pero su efecto práctico es que el conjunto de restricciones publicado no es el conjunto de restricciones operativo para cada implementación. Una institución que contrata una implementación configurada no puede saber a partir de este documento por sí solo qué versión del margen de comportamiento del modelo está adquiriendo, y el documento no ofrece ningún mecanismo de divulgación que le permita preguntarlo.

El trabajo de evaluación conlleva una tercera limitación, más inmediata. Microsoft afirma que sus modelos aún no están entrenados con el Código de Conducta, que el programa de evaluación está en desarrollo, y que los escenarios ilustrativos son sintéticos, conversacionales en lugar de agénticos o multimodales, y generados usando uno de sus propios modelos, con los ejemplos desalineados producidos indicando a ese modelo que fallara de maneras específicas. Esta es una forma razonable de arrancar un conjunto de evaluación y una base irrazonable para la garantía institucional, porque las propias generaciones de un modelo de comportamiento bueno y malo codifican los mismos supuestos con los que fue entrenado el modelo. Nada en el apéndice constituye evidencia de que un sistema implementado se comporte según lo especificado, y Microsoft no afirma que lo haga. Las pruebas adversariales independientes, la revisión humana experta por especialistas del dominio y la medición longitudinal en interacciones reales se reconocen todas como trabajo futuro. Hasta que ese trabajo exista y se publique, la postura responsable en materia de contratación es que las afirmaciones sobre comportamiento en este documento no están validadas.

La brecha de dominio para los sistemas de lengua de señas

Para las instituciones que implementan IA de lengua de señas, el problema más consecuente es que el documento es agnóstico en cuanto al dominio por diseño, y por lo tanto los daños específicos de la tecnología de lengua de señas no aparecen en él en ningún lugar. Las restricciones absolutas abordan armas, ciberoperaciones, pérdida de control, manipulación a gran escala, respuesta a crisis, suplantación de identidad, seguridad infantil, discriminación, contenido explícito y violencia. Nada en esa lista, y nada en las directrices operativas que la siguen, alcanza los modos de falla que real-

mente determinan si un sistema de lengua de señas es seguro para ponerlo frente a personas Sordas y Sordociegas.

La validez lingüística es la primera de ellas. Un sistema de generación de lengua de señas puede producir una salida que sea fluida en apariencia e inválida en sustancia, violando la estructura gramatical de la lengua de destino, usando incorrectamente marcadores no manuales que llevan peso sintáctico, colapsando distinciones de registro, o transformando la variación regional y comunitaria en un estándar sintético que ninguna comunidad usa en realidad. Las disposiciones de precisión del Código de Conducta abordan la corrección factual, el sourcing y la confianza calibrada, ninguna de las cuales detecta un enunciado señado gramaticalmente mal formado. Un sistema puede satisfacer todos los requisitos de honestidad de la Parte 3 mientras produce un lenguaje que un señante fluido reconocería de inmediato como incorrecto, y el marco no contiene ningún instrumento que haga visible el fallo.

Los daños que se derivan de ese fallo están igualmente ausentes. La implementación en entornos educativos expone a niños en proceso de adquisición del lenguaje a entradas mal formadas, con consecuencias para el desarrollo de la fluidez que ninguna categoría general de seguridad capta. La implementación en entornos profesionales y cívicos interpretados puede tergiversar el discurso de una persona Sorda ante una audiencia oyente, con consecuencias reputacionales y legales que recaen enteramente sobre la persona cuyo lenguaje fue alterado. La implementación en contextos de accesibilidad pública puede generar la apariencia de acceso sin entregar ninguno, lo cual es un daño de cumplimiento tanto como lingüístico. Estos son resultados ordinarios y previsibles de la tecnología, y un documento de gobernanza que no los nombra no puede ser el documento de gobernanza para ella.

Otras dos brechas son estructurales más que lingüísticas. El documento organiza la autoridad alrededor de Microsoft, los operadores y los usuarios, y no tiene ninguna categoría para la comunidad lingüística cuyo idioma reproduce el sistema. Para los sistemas de lengua de señas, la autoridad comunitaria no es una cortesía ética añadida sobre la validación técnica; es la base misma de la validación, porque los señantes fluidos son la única fuente de verdad sobre si el resultado es aceptable. Un marco que solo reconoce a la institución adquirente y al usuario final no tiene lugar donde colocar esa autoridad ni mecanismo para exigirla. En un sentido relacionado, el documento trata la accesibilidad como un tema que el modelo debe manejar bien en lugar de una propiedad que el sistema mismo debe tener. Si la interfaz, los mecanismos de su-

pervisión, la divulgación del estatus de IA y las vías de escalamiento son en sí mismos utilizables por personas Sordas y Sordociegas es una pregunta que el texto nunca formula, y un sistema de lengua de señas que falla en ese punto ha fallado en el punto mismo del ejercicio.

Cuestiones estructurales que el documento deja abiertas

Más allá de la brecha de dominio, tres cuestiones estructurales importarán a cualquiera que dependa de este marco. La primera concierne al pluralismo. Microsoft se compromete a respaldar una amplia gama de visiones del mundo y afirma que el pluralismo no significa neutralidad ante el daño, lo cual es la matización correcta. Sin embargo, el marco descansa sobre una posición de valores específica e identificable, centrada en la autonomía individual, el consentimiento informado, la agencia personal y una concepción de la plenitud arraigada en las tradiciones liberales de derechos, y presenta esa posición como la base inclusiva dentro de la cual operan los valores plurales. No hay nada de malo en construir un marco de gobernanza sobre valores declarados. La dificultad es que declarar el pluralismo mientras se operacionaliza una sola tradición oculta la elección a las instituciones que la heredan, y algunas de esas instituciones sirven a comunidades cuyas normas de gobernanza, particularmente en torno a la autoridad colectiva sobre el patrimonio cultural y lingüístico compartido, no se reducen limpiamente a la preferencia individual del usuario.

La segunda concierne a la rendición de cuentas. Se dice que los operadores asumen la responsabilidad de sus configuraciones y usos, pero el documento no especifica ninguna obligación de divulgación, ningún derecho de auditoría, ni ningún mecanismo de adjudicación para las disputas entre un operador y las personas a las que afecta su implementación. La responsabilidad afirmada sin una vía procesal para hacerla cumplir es una declaración de intención. Las instituciones que dependan de este documento deben asumir que cualquier rendición de cuentas que necesiten tendrán que construirla en sus propios contratos.

El tercero concierne a la resolución de conflictos. El marco acepta que sus objetivos a veces competirán entre sí, y dirige a los modelos a consultar las directrices operativas, declarar límites al usuario, proponer alternativas y, cuando nada más resuelve la tensión, actuar según una lectura holística del documento en su conjunto. Como filosofía de diseño para un modelo bajo incertidumbre, eso es defendible y probablemente inevitable. Como especificación, significa que la resolución de conflictos de valores ge-

nuinos se delega al juicio del modelo y no es predecible de antemano ni revisable después. Para interacciones de bajo riesgo esto es aceptable. Para un sistema que media el acceso de una persona Sorda a la atención médica, el empleo o el proceso legal, no lo es.

Lo que esto significa para las adquisiciones

Una institución que adquiere IA de lengua de señas puede usar este documento como arquitectura de referencia, y no debería usarlo como especificación. La cadena de mando, el tratamiento bidireccional de la falla de seguridad y la descomposición de los comportamientos en subcomportamientos verificables valen la pena de adoptarse. Lo que hay que añadir es todo aquello que el marco general no puede proporcionar.

Esa incorporación comienza con la validación lingüística. El lenguaje de contratación debe especificar la lengua y variedades de destino, definir la salida aceptable a nivel de estructura gramatical y registro, exigir la evaluación por señantes Sordos fluidos en lugar de métricas indirectas o revisores oyentes, y establecer la frecuencia y el método con que se medirá la calidad lingüística después de la implementación, no solo antes de ella. La autoridad comunitaria debe incorporarse al instrumento como un rol permanente con alcance definido, no mencionarse vagamente como participación de partes interesadas, y debe conllevar la capacidad de rechazar la salida y detener la implementación en lugar de simplemente comentar sobre ella. El daño debe operacionalizarse en términos de dominio, convirtiendo las variables de proporcionalidad de Microsoft en categorías de fallo con nombre y umbrales que un auditor pueda aplicar. La evidencia de evaluación debe exigirse antes de la implementación y debe ser evidencia sobre el sistema configurado en el caso de uso previsto, no afirmaciones generales de comportamiento del proveedor. La resolución de disputas debe especificarse contractualmente, incluyendo qué ocurre cuando una configuración del operador entra en conflicto con el juicio comunitario sobre la aceptabilidad lingüística. La accesibilidad del propio sistema, incluidos sus mecanismos de supervisión y escalamiento, debe ser una condición de aprobación o rechazo, no una aspiración de diseño.

Conclusión

Microsoft ha producido una declaración seria e inusualmente transparente de su filosofía de gobernanza, y ha invitado a la crítica de la misma durante una ventana de consulta definida en lugar de presentarla como algo ya resuelto. Tiene éxito como de-

claración de principios y de la estructura de autoridad que esos principios requieren. Aún no proporciona la precisión operativa, la evidencia validada ni la taxonomía de daños específica del dominio de la que depende la adquisición institucional, y no pretende hacerlo.

El trabajo que queda es el trabajo que siempre ha quedado pendiente: traducir los principios de gobernanza en instrumentos que los responsables de adquisiciones puedan incorporar a los contratos, que los auditores puedan aplicar después de la implementación, y que las comunidades afectadas puedan invocar cuando un sistema les falle. Para la IA de lengua de señas, esa traducción no puede heredarse de un marco general. Tiene que ser construida por las personas que saben cómo se ve la falla en el idioma.

> **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.

Cómo 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