

Chapter 7 / Capítulo 7

AI, metaverse and sustainability in the new
organizational design
ISBN: 978-9915-9928-8-4
DOI: 10.62486/978-9915-9928-8-4.Ch07
Pages: 43-48



© 2026 The authors. This is an open access
book chapter distributed under the terms of
the Creative Commons Attribution (CC BY)
4.0 License.

Artificial intelligence ethics and organizational responsibility

Ética de la inteligencia artificial y responsabilidad organizacional

Jesús Alejandro Vázquez Meza¹ 

Víctor Manuel Martínez García² 

Yennifer Díaz Romero³ 

Manuel Iván Tostado Ramírez⁴ 

Rogelio Estrada Lizárraga⁴ 

José de Jesús Reyes Machain² 

1 Universidad Autónoma de Sinaloa, Unidad Académica Preparatoria Rubén Jaramillo, Vicerrectoría, Sinaloa, México

2 Universidad Autónoma de Sinaloa, Facultad de Arquitectura y Diseño, Vicerrectoría, Sinaloa, México

3 Universidad Autónoma de Sinaloa, Facultad de Ciencias de la Educación, Vicerrectoría, Sinaloa, México

4 Universidad Autónoma de Sinaloa, Facultad de Informática, Vicerrectoría, Sinaloa, México

5 Universidad Autónoma de Sinaloa, Facultad de Informática, Vicerrectoría, Sinaloa, México

ABSTRACT

This chapter reviews ethical issues that arise when artificial intelligence influences organizational processes and decisions. The discussion is organized around transparency, accountability, fairness, non-discrimination, and algorithmic bias. It examines the importance of enabling people to understand the scope and limits of AI systems, challenge their outputs, and maintain meaningful oversight. Institutional responsibilities are connected with the privacy of data used or processed by AI applications. Employee well-being is also considered, including possible effects on autonomy, safety, and the experience of work. Ethical principles are not treated as sufficient declarations in themselves; they require procedures, identifiable responsibilities, and controls. The review distinguishes technical performance from the social conditions in which a tool is used. The chapter argues that responsible integration requires clarity about who makes decisions, who reviews them, and who responds to adverse consequences. Data practices and human participation are therefore integral to the ethical assessment of the design and operation of these systems, rather than additional issues to consider only after deployment.

Keywords: AI ethics; Algorithmic bias; Transparency; Privacy; Organizational accountability

RESUMEN

El capítulo revisa los problemas éticos que surgen cuando la inteligencia artificial interviene en procesos y decisiones organizacionales. Ordena la discusión en torno a la transparencia, la responsabilidad, la equidad, la no discriminación y los sesgos algorítmicos. Examina la importancia de que las personas puedan comprender los alcances y límites de los sistemas, cuestionar sus resultados y mantener formas significativas de supervisión. El texto relaciona las obligaciones institucionales con la privacidad de los datos que alimentan o atraviesan las aplicaciones de inteligencia artificial. También incorpora el bienestar de los trabajadores como criterio de evaluación, considerando posibles efectos sobre su autonomía, seguridad y experiencia laboral. Los principios éticos no se presentan como declaraciones suficientes por sí mismas, sino como orientaciones que requieren procedimientos, responsables y controles. La revisión diferencia el desempeño técnico de las condiciones sociales bajo las cuales se utiliza una herramienta. El capítulo sostiene que integrar inteligencia artificial de manera responsable exige identificar quién decide, quién revisa y quién responde ante consecuencias adversas. Asimismo, muestra que el tratamiento de los datos y la participación humana forman parte de la evaluación ética del diseño y la operación de estos sistemas. La exposición examina estas exigencias de manera interdependiente: una práctica transparente puede seguir siendo problemática si reproduce sesgos, y una herramienta precisa puede resultar inadecuada si las personas afectadas no disponen de mecanismos de revisión.

Palabras clave: Ética de la inteligencia artificial; Sesgo algorítmico; Transparencia; Privacidad; Responsabilidad organizacional

7.1 TRANSPARENCIA

Cuando un sistema de inteligencia artificial toma, o influye en, una decisión que afecta a una persona, esa persona tiene derecho a entender, al menos en términos generales, cómo se llegó a esa decisión. Aunque el principio parece de sentido común, en ocasiones resulta difícil de garantizar en la práctica, sobre todo cuando el sistema en cuestión opera con modelos cuya lógica interna ni siquiera sus propios desarrolladores pueden explicar del todo.

En su *Recomendación sobre la Ética de la Inteligencia Artificial*, la UNESCO (2021) establece la transparencia como uno de los principios rectores de cualquier desarrollo o uso responsable de estas tecnologías, junto con la rendición de cuentas, la inclusión y la sostenibilidad. De acuerdo con el documento, la transparencia consiste en garantizar que exista suficiente claridad sobre cómo y cuándo se utiliza la IA, de manera que las personas afectadas puedan comprender sus resultados y, cuando corresponda, cuestionarlos, sin necesidad de exponer aspectos técnicos que probablemente para la mayoría de las personas pasarían desapercibidos.

Barroso Camiade y Pérez Castrejón (2025) sitúan este principio dentro de un debate más amplio sobre los supuestos filosóficos de la propia inteligencia artificial, esto al retomar la crítica clásica de Hubert Dreyfus a las asunciones sobre las que se construyeron los primeros sistemas inteligentes. Su argumento, actualizado al contexto actual, plantea que la opacidad de muchos sistemas de IA, además de ser un problema técnico, es una consecuencia directa de haber construido la tecnología sobre supuestos que sus propios diseñadores nunca sometieron a suficiente escrutinio, de modo que, entendida así, la transparencia deja de ser únicamente una buena práctica de comunicación para convertirse también en una exigencia epistemológica sobre qué tan bien se entiende lo que estos sistemas hacen.

Los datos del *AI Index Report* de la Universidad de Stanford (2025) muestran que esta exigencia sigue lejos de cumplirse en la práctica. El índice de transparencia entre los principales desarrolladores de modelos de frontera pasó de un 37 % en octubre de 2023 a un 58 % en mayo de 2024, lo cual representa una mejora real, aunque la mayoría de estos sistemas todavía está lejos de un nivel de divulgación que permita una auditoría externa sobre sus datos de entrenamiento, su consumo de cómputo y su impacto posterior al despliegue. El mismo informe menciona, además, un aumento en los incidentes de IA reportados formalmente, los cuales llegaron a 233 en 2024, lo que representa un aumento de más del 56 % con respecto al año anterior (Stanford University, 2025). La brecha entre la velocidad con la que se despliegan estos sistemas y la velocidad con la que se documentan sus fallas es, en sí misma, uno de los indicios más claros de por qué la transparencia sigue siendo un principio aspiracional, más que una práctica consolidada.

7.2 RESPONSABILIDAD

El determinar quién responde cuando un sistema de inteligencia artificial se equivoca es uno de los vacíos legales y organizacionales más persistentes en este terreno. Al respecto, Méndez Reyes (2025) documenta que, conforme la IA se expande hacia *chatbots*, redes neuronales y conducción autónoma, entre muchas otras aplicaciones, trae consigo problemas de sesgo, opacidad algorítmica y preocupaciones sobre la privacidad y la seguridad de los datos, mismos que hacen indispensable establecer marcos claros de responsabilidad organizacional.

Por su parte, la UNESCO (2021) plantea que la responsabilidad debe distribuirse de forma proporcional entre quienes diseñan, despliegan y utilizan estos sistemas, con lo cual se evita tanto el extremo de exonerar por completo a la organización que decide implementarlos como el de atribuir toda la responsabilidad a quien programó el algoritmo original. En la práctica, esto significa que ninguna organización puede refugiarse en el argumento de que “el sistema decidió así” para eludir su propia responsabilidad sobre las consecuencias de haber elegido implementar, sin una supervisión adecuada, una herramienta capaz de afectar a personas reales.

El *AI Index Report* de Stanford (2025) documenta que las organizaciones están empezando a internalizar esta exigencia, aunque de forma todavía desigual, y señala que la proporción de empresas que declara no tener ninguna política de inteligencia artificial responsable cayó de forma notable, de 24 a 11 % en apenas un año. Se trata de una señal alentadora; no obstante, el mismo informe advierte que reconocer los riesgos y mitigarlos activamente son cosas distintas. Tal es el caso de una encuesta de McKinsey citada en el índice, la cual muestra que, aunque la mayoría de las organizaciones identifica correctamente los riesgos centrales de la IA responsable, no todas traducen ese reconocimiento en acciones concretas de mitigación. Dicho de otra forma, saber que se es responsable todavía no equivale a actuar en consecuencia.

Barroso Camiade y Pérez Castrejón (2025) añaden un matiz filosófico importante a esta discusión al retomar la distinción clásica entre ética descriptiva y ética normativa, a partir de la cual señalan que buena parte del debate público sobre la responsabilidad de la IA se queda atrapado en describir lo que distintos grupos consideran correcto o incorrecto, sin avanzar hacia principios normativos claros sobre cómo debería actuar una organización en la práctica. Dicha distinción trasciende lo académico, ya que explica por qué tantas organizaciones logran redactar códigos de ética extensos y bien intencionados que, sin embargo, no se traducen en ningún cambio observable en la manera en que se toman las decisiones. Una responsabilidad genuina exige cruzar ese puente entre la descripción de valores y la normatividad operativa, algo que la mayoría de los marcos éticos corporativos todavía no logra del todo.

7.3 EQUIDAD Y NO DISCRIMINACIÓN

Un sistema de inteligencia artificial entrenado con datos que reflejan sesgos históricos, como los de género, origen étnico o clase socioeconómica, tiende, salvo que haya una intervención deliberada, a reproducir y en ocasiones a amplificar esos mismos sesgos en sus resultados. Esto ocurre sin que el sistema sea “malintencionado” en ningún sentido, puesto que aprendió lo que los datos le enseñaron, incluidas sus injusticias.

Guaña-Moya y Chipuxi-Fajardo (2023) analizan este fenómeno desde la intersección entre la ética y la privacidad de los datos, en donde señalan que garantizar la equidad requiere algo más que buenas intenciones declaradas, esto es, principios operativos concretos, como lo son la transparencia, la responsabilidad y el respeto a los derechos fundamentales de las personas, principios que tienen que traducirse en procesos verificables, más allá de las declaraciones de valores corporativos. La UNESCO (2021) refuerza esta idea de que la inclusión sea uno de sus principios centrales, misma que, lejos de ser un gesto simbólico, se entiende como la exigencia activa de que los sistemas de IA no perpetúen ni profundicen las desigualdades existentes.

El *Future of Jobs Report 2025* aporta a esta discusión una dimensión adicional, poco tratada fuera del ámbito laboral, según la cual la equidad algorítmica importa dentro de un sistema de IA ya desplegado, pero también en quién tiene acceso, desde el principio, a formarse en las competencias que esa IA exige. De acuerdo con el informe, el 85 % de los empleadores planea invertir en capacitación en la capacitación de su personal de forma general, en lo que para la capacitación ligada a la IA es del 77% (World Economic Forum, 2025); sin embargo, si esa misma inversión no se distribuye de forma deliberadamente equitativa entre distintos niveles jerárquicos, géneros y regiones dentro de una organización, corre el riesgo de convertirse en un mecanismo que amplía, en lugar de reducir, las brechas existentes entre quienes ya tenían acceso a oportunidades de desarrollo y quienes no. En ese sentido, la equidad algorítmica empieza antes de que un sistema tome su primera decisión, es decir, en quién tuvo la oportunidad de aprender a trabajar con él.

7.4 SESGOS ALGORÍTMICOS

Conviene distinguir el sesgo algorítmico, como fenómeno técnico específico, de la discusión más amplia sobre la equidad. Este sesgo ocurre cuando un modelo produce resultados sistemáticamente distintos a los esperados, y que por lo general resultan perjudiciales, para determinados grupos de personas; en ocasiones quienes diseñaron o implementaron el sistema no se percataron de esta situación en tiempo y forma.

Méndez Reyes (2025) identifica los sesgos algorítmicos como uno de los problemas centrales que la inteligencia artificial trae consigo en su expansión reciente, junto con la falta de transparencia y las preocupaciones sobre la privacidad. Su análisis sugiere que estos sesgos rara vez son producto de una decisión de diseño explícita, ya que suelen colarse a través de los datos de entrenamiento, de variables *proxy* aparentemente neutrales que en realidad se correlacionan con características protegidas o de la falta de diversidad entre quienes diseñan y prueban el sistema antes de su despliegue. Para detectarlos se requieren auditorías activas y deliberadas, lo cual va más allá de la simple confianza en que “el sistema funcionará bien porque fue diseñado con buena intención”.

La evidencia empírica más reciente muestra que ni siquiera los esfuerzos explícitos por eliminar el sesgo garantizan resultados neutrales. El *AI Index Report* de Stanford (2025) documenta que varios de los modelos de lenguaje más avanzados, incluidos sistemas diseñados

de forma deliberada con medidas para reducir sesgos explícitos, siguen exhibiendo sesgos implícitos considerables. Estos modelos asocian con más frecuencia términos negativos con personas afrodescendientes, vinculan a las mujeres con las humanidades en lugar de con disciplinas científicas y tecnológicas y favorecen a los hombres para roles de liderazgo, con lo cual refuerzan patrones discriminatorios de género y raza en la toma de decisiones. Si bien el hallazgo resulta incómodo, es necesario reconocer que mejorar el desempeño de un sistema en pruebas estandarizadas de sesgo no equivale a haber resuelto el problema de fondo, y que una organización que confía ciegamente en esas pruebas puede seguir desplegando decisiones sesgadas sin saberlo.

7.5 SUPERVISIÓN HUMANA

Ninguno de los principios éticos revisados hasta aquí, como lo son la transparencia, la responsabilidad o la equidad, se sostiene sin un mecanismo concreto que lo haga cumplir en la práctica, y ese mecanismo es, casi invariablemente, la supervisión humana activa.

Autio et al. (2024), en el perfil complementario de este mismo marco dedicado a la inteligencia artificial generativa, subrayan que los riesgos de estos sistemas cambian constantemente conforme evolucionan los modelos, por lo que la supervisión humana tiene que concebirse como un proceso continuo que se ajuste a la misma velocidad que la propia tecnología, en lugar de una revisión puntual al momento de la implementación. Una organización que instale un sistema de IA y considere terminado el trabajo de supervisión ha malentendido, de raíz, en qué consiste la responsabilidad ética sobre estas herramientas.

El estudio experimental de Dell'Acqua et al. (2023) ofrece evidencia concreta de por qué esta supervisión continua, más que un formalismo burocrático, es una necesidad práctica medible. En la tarea diseñada para exigir un juicio contextual complejo, fuera de lo que los autores llaman la frontera de capacidad del sistema, los consultores que confiaron en la IA sin una supervisión crítica vieron caer su precisión de forma marcada frente a quienes trabajaron sin ella. A partir de ello se puede ver que la falla estuvo menos en la tecnología que en la ausencia de un mecanismo que alertara a los consultores sobre cuándo debían aplicar un mayor escrutinio a las sugerencias del sistema. Se trata, en esencia, de la misma advertencia que hacen Tabassi y Autio et al. desde el terreno normativo, en el sentido de que, sin una supervisión humana activa y constante, ni el sistema más sofisticado garantiza decisiones correctas.

7.6 PRIVACIDAD

Para funcionar bien, los sistemas de inteligencia artificial necesitan datos, a menudo muchos, y con frecuencia datos personales sensibles. Esa dependencia estructural pone a la privacidad en tensión directa con la utilidad misma de estas herramientas, por lo que resolver esa tensión de forma responsable es uno de los retos éticos más persistentes del sector.

Guaña-Moya y Chipuxi-Fajardo (2023) concluyen su análisis con un llamado directo, el cual consiste en que garantizar un desarrollo responsable y respetuoso de la inteligencia artificial exige tanto reflexión ética como una regulación legal adecuada, lo que incluye la adopción de principios éticos sólidos, la promoción activa de la transparencia y la protección explícita de la privacidad de los datos de las personas. En la misma línea, la UNESCO (2021) coincide en que ningún desarrollo de IA puede considerarse ético si compromete de manera sistemática el derecho de las personas a controlar qué se sabe de ellas y cómo se usa esa información.

7.7 BIENESTAR DE LOS TRABAJADORES

Este capítulo cierra con un principio que rara vez encabeza las listas de prioridades éticas sobre la IA, pero que debería estar mucho más presente, el cual tiene que ver con el impacto de estos sistemas en la salud mental, la seguridad laboral y el sentido de propósito de las personas que trabajan junto a ellos o bajo su supervisión algorítmica.

Méndez Reyes (2025) sitúa esta preocupación dentro de un marco de teoría del discurso, esto al argumentar que la manera en que las organizaciones hablan y deciden sobre la inteligencia artificial tiene consecuencias directas en la forma en que las personas experimentan su propio trabajo cuando esa tecnología entra en juego. Un sistema de IA que optimiza la productividad sin ninguna consideración sobre el bienestar de quienes trabajan con él puede generar ganancias

de corto plazo mientras erosiona, de forma silenciosa y acumulativa, la confianza, la motivación y la salud de la fuerza laboral. En última instancia, más que un ejercicio abstracto de principios bien redactados, la ética de la inteligencia artificial es la pregunta concreta de si las personas que trabajan dentro de una organización terminan mejor o peor como consecuencia de las decisiones que esa organización tomó sobre la manera de usar esta tecnología.

Los datos de Deloitte (2025) muestran que esta preocupación no se limita a los especialistas en ética y que ya es ampliamente compartida por la propia fuerza laboral, tal como se observa en su informe *Global Human Capital Trends*, en el cual se encontró que más de la mitad de los trabajadores y líderes encuestados en 93 países expresa preocupación por la creciente ambigüedad entre el trabajo que realiza una persona y el que realiza la inteligencia artificial. Si no se atiende de forma deliberada, esa incertidumbre puede convertirse con facilidad en ansiedad laboral sostenida. Es así como las organizaciones que ignoran esta señal, al tratar el bienestar como un asunto separado de su estrategia de IA en lugar de considerarlo una dimensión inseparable de ella, terminan pagando ese costo más tarde en forma de rotación, desmotivación y pérdida de confianza institucional.