Antes del algoritmo: ética de la información y responsabilidad en los datos de entrenamiento de la inteligencia artificial generativa

Before the Algorithm: Information Ethics and Responsibility in Generative Artificial Intelligence Training Data

Alejandro Villegas Muro
avillegas@uach.mx

ORCID: https://orcid.org/0000-0001-6362-5137

Universidad Autónoma de Chihuahua

DOI: https://doi.org/10.54167/roee.v4i2.2446

Recibido: 2026/08/21

Aceptado para su publicación: 2026/08/25

Publicado: 2026/08/31

Resumen: La inteligencia artificial generativa utiliza grandes volúmenes de información para entrenar modelos computacionales. Este artículo sostiene que la responsabilidad ética debe examinarse desde una etapa previa a los resultados: la selección, recopilación, transformación e incorporación de información a los conjuntos de entrenamiento. Mediante una investigación documental crítico-interpretativa se analizan aportaciones sobre ética de la información, documentación de conjuntos de datos, procedencia, gobernanza, sesgos, consentimiento y marcos internacionales de inteligencia artificial. A partir del análisis se propone una cadena de responsabilidad informacional con seis dimensiones: procedencia, legitimidad de uso, calidad, representación, trazabilidad y responsabilidad. La propuesta distingue accesibilidad técnica de legitimidad ética y plantea que una inteligencia artificial responsable requiere gobernar también las condiciones informacionales que hacen posible su desarrollo tecnológico.
Palabras clave: inteligencia artificial, ética de la tecnología, datos de entrenamiento, procedencia de la información, responsabilidad informacional.

Abstract: Generative artificial intelligence uses large volumes of information to train computational models. This article argues that ethical responsibility should be examined before model outputs, beginning with the selection, collection, transformation, and incorporation of information into training datasets. Through critical-interpretive documentary research, it analyzes specialized contributions on information ethics, dataset documentation, provenance, governance, bias, consent, and recent international artificial intelligence frameworks. Based on this analysis, the article proposes an information responsibility chain composed of six dimensions: provenance, legitimacy of use, quality, representation, traceability, and responsibility. The proposal distinguishes technical accessibility from ethical legitimacy and argues that responsible artificial intelligence requires governance not only of models and outputs, but also of the informational conditions, decisions, and practices that make their technological development possible.
Keywords: artificial intelligence, ethics of technology, training data, information provenance, information responsibility.

I. Introducción

Antes de que un gran modelo de lenguaje genere contenidos, necesita haber sido entrenado con grandes cantidades de datos. Buena parte procede de información creada originalmente con otros propósitos: artículos científicos, libros, páginas institucionales, conversaciones, publicaciones en redes sociales, código informático y materiales depositados en repositorios.

Cuando estos contenidos se recopilan, filtran, fragmentan, etiquetan, combinan o convierten en unidades procesables para un sistema, ocurre una transformación que no es solamente técnica. También cambia su contexto de utilización.

Esta cuestión se vuelve especialmente importante en los modelos generativos debido a la escala de los conjuntos utilizados para su entrenamiento. La investigación sobre grandes corpus de texto ha mostrado que las fuentes recopiladas mediante rastreo de la web pueden ser heterogéneas, insuficientemente documentadas y resultado de múltiples procesos de

filtrado. El estudio de Jesse Dodge y colaboradores sobre el corpus C4 identificó procedencias inesperadas y mostró que determinadas prácticas de filtrado modificaban la composición del conjunto de datos.[1] Su análisis resulta importante porque evidencia que un corpus aparentemente técnico incorpora decisiones sobre qué información entra, cuál se elimina y qué termina representando el mundo textual disponible para el modelo.

El problema tampoco se limita a la calidad técnica. Emily M. Bender y Batya Friedman habían planteado, antes de la actual expansión de la inteligencia artificial generativa (IAG), la necesidad de documentar las poblaciones y contextos de los datos utilizados en el procesamiento del lenguaje natural.[2] Su argumento partía de una preocupación científica y ética: emplear información producida por determinados grupos para desarrollar sistemas destinados a otros puede generar exclusiones, errores de generalización y representaciones inadecuadas. Más adelante, Timnit Gebru y colaboradores propusieron las *datasheets for datasets* como una práctica sistemática para documentar la motivación, composición, recopilación y usos previstos de los conjuntos de datos.[3] La idea común en estos trabajos es sencilla pero relevante: comprender un sistema requiere conocer también aquello con lo que fue construido.

La aparición de modelos de propósito general ha aumentado la dificultad. En muchos casos ya no se trata de un conjunto pequeño diseñado para una tarea concreta, sino de agregaciones de fuentes distintas que atraviesan repositorios, conjuntos derivados, procesos de limpieza y nuevas combinaciones. Shayne Longpre y colaboradores, mediante una auditoría de más de 1,800 conjuntos de datos textuales, encontraron problemas importantes en la documentación de licencias y atribución.[4] El problema de la procedencia adquiere así una dimensión práctica. Cuando el linaje de los datos se vuelve difícil de reconstruir, también se dificulta determinar qué autorizaciones existían, cuáles eran las condiciones originales de uso y quién debería responder por decisiones tomadas a lo largo de la cadena.

A esto se añade el consentimiento. La web fue construida para facilitar publicación, enlace, lectura, indexación y circulación de información, no para resolver de manera anticipada todos los usos que varias décadas después serían posibles mediante sistemas generativos. Un análisis posterior de Longpre y colaboradores sobre miles de dominios web encontró un incremento de restricciones específicas relacionadas con la utilización de contenidos por sistemas de inteligencia artificial y, al mismo tiempo, inconsistencias entre diferentes mecanismos empleados para expresar preferencias de acceso y reutilización.[5] El fenómeno muestra que la discusión no puede resolv-

1 Jesse Dodge et al., "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," en Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. (Association for Computational Linguistics, 2021), 1286–1305, https://doi.org/10.18653/v1/2021.emnlp-main.98.

2 Emily M. Bender y Batya Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," Transactions of the Association for Computational Linguistics 6 (2018): 587–604, https://doi.org/10.1162/tacl_a_00041.

3 Timnit Gebru et al., "Datasheets for Datasets," Communications of the ACM 64, núm. 12 (2021): 86–92, https://doi.org/10.1145/3458723.

4 Shayne Longpre et al., "A Large-Scale Audit of Dataset Licensing and Attribution in AI," Nature Machine Intelligence 6 (2024): 975–87, https://doi.org/10.1038/s42256-024-00878-8.

5 Shayne Longpre et al., "Consent in Crisis: The Rapid Decline of the AI Data Commons," Advances in Neural Information Processing Systems 37 (2024): 108042–87, https://doi.org/10.52202/079017-3431.

erse simplemente identificando qué contenido era técnicamente descargable.

La preocupación también está presente en los marcos internacionales. La *Recomendación sobre la Ética de la Inteligencia Artificial* de la UNESCO plantea que la gobernanza debe atender la calidad de los datos de entrenamiento, los procesos de recopilación y selección, la privacidad, la diversidad y la responsabilidad a lo largo del ciclo de vida del sistema.[6] La preocupación se ha trasladado asimismo hacia instrumentos regulatorios. El *Reglamento de Inteligencia Artificial* de la Unión Europea incorpora obligaciones para los proveedores de determinados modelos de propósito general relacionadas con documentación, derechos de autor e información sobre el contenido utilizado durante su entrenamiento.[7]

Estos antecedentes permiten formular el problema que guía este trabajo: ¿qué principios éticos deberían regular la selección, incorporación y utilización de información como datos de entrenamiento para sistemas de IAG?

La hipótesis es que la ética de la inteligencia artificial resulta incompleta cuando se concentra principalmente en el funcionamiento, despliegue y resultados de los modelos. Una inteligencia artificial responsable requiere también una ética de la información de entrada, capaz de considerar procedencia, legitimidad de uso, calidad, representación, trazabilidad y responsabilidad durante la construcción de los conjuntos empleados para entrenar los sistemas. Dicho de otra forma, que una información sea técnicamente accesible no significa necesariamente que cualquier forma de incorporación, reutilización o transformación resulte éticamente equivalente.

El objetivo es desarrollar este argumento y proponer, como resultado del análisis, una cadena de responsabilidad informacional para los datos de entrenamiento. Esta cadena comprende seis dimensiones relacionadas entre sí: 1) procedencia, 2) legitimidad de uso, 3) calidad informativa, 4) representación, 5) trazabilidad y 6) responsabilidad. No pretende sustituir los marcos existentes de gobernanza de datos ni presentarse como una lista exhaustiva de requisitos. Su finalidad es ofrecer un esquema conceptual que permita observar el proceso de entrenamiento desde la perspectiva de las condiciones informacionales que lo hacen posible.

El trabajo se desarrolla mediante una investigación documental de alcance teórico y orientación crítico-interpretativa. El corpus de análisis estuvo integrado por 17 documentos: 14 aportaciones académicas y tres documentos normativos o institucionales, seleccionados por su relación directa con la ética de la información, la documentación de conjuntos de datos, los sesgos en corpus de gran escala, la procedencia, el consentimiento, la transparencia y la gobernanza de la inteligencia artificial. Estos documentos permitieron identificar la forma en que dichos problemas han comenzado a incorporarse a la regulación y a las políticas de inteligencia artificial.

II. De la ética de la información a la ética de los datos de entrenamiento

La ética de la información ofrece un punto de partida adecuado porque permite ampliar el análisis más allá del dispositivo o del algoritmo concreto. Luciano Floridi ha desarrollado este campo como una perspectiva capaz de estudiar problemas morales relaciona-

6 UNESCO, Recommendation on the Ethics of Artificial Intelligence (Paris: UNESCO, 2021).
7 European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2024.

dos con la creación, procesamiento, distribución y utilización de información dentro de una sociedad profundamente mediada por tecnologías digitales.[8] Su propuesta supera una comprensión estrictamente instrumental de las tecnologías de información. La información forma parte de un entorno en el que las personas actúan, toman decisiones, construyen identidades y establecen relaciones.

Conviene distinguir tres conceptos que suelen emplearse de forma intercambiable: información, dato y dato de entrenamiento. Aunque sus fronteras pueden variar según la disciplina, la distinción permite observar la transformación que interesa en este trabajo.

La información se entiende aquí como contenido producido, organizado o comunicado dentro de un contexto que le proporciona significado. Una noticia, un artículo científico, una fotografía, una conversación o una obra literaria no constituyen simplemente secuencias de caracteres o píxeles. Tienen autoría o procedencia, público, finalidad, momento de producción, condiciones de acceso y relaciones con otros contenidos.

El dato es una representación que permite registrar, almacenar, relacionar o procesar determinados elementos de esa información. La conversión a datos facilita operaciones computacionales, pero puede reducir o separar parte del contexto original.

Cuando un documento es fragmentado en unidades para procesamiento, por ejemplo, la estructura computacional obtenida no conserva necesariamente todas las relaciones sociales o editoriales que acompañaban al documento.

El dato de entrenamiento añade otra transformación: esos registros son utilizados para ajustar parámetros, aprender regularidades, construir representaciones o modificar el comportamiento de un sistema. Desde el punto de vista computacional esta transformación es necesaria. Desde el punto de vista ético, sin embargo, importa reconocer que el cambio de función no elimina de manera automática las preguntas asociadas con la procedencia.

Por ello, la ética de los datos de entrenamiento no puede reducirse al cumplimiento de requisitos mínimos de formato, limpieza o eficiencia. Las decisiones sobre datos tienen consecuencias sobre el sistema. Nithya Sambasivan y colaboradores denominaron data cascades a los efectos acumulativos que problemas surgidos en el trabajo con datos pueden producir en fases posteriores de proyectos de inteligencia artificial.[9] Su estudio mostró algo que suele perderse cuando el modelo ocupa el centro de la atención: las prácticas relacionadas con los datos son fundamentales, aunque dentro de ciertos procesos de desarrollo hayan recibido menor reconocimiento que el trabajo sobre modelos.

Bender y colaboradores han señalado que los grandes modelos lingüísticos presentan riesgos que no pueden comprenderse sin atender a los conjuntos masivos utilizados para entrenarlos.[10] Un corpus obtenido de internet no constituye una representación neutral de la humanidad. Refleja quién tiene acceso

8 Luciano Floridi, "Information Ethics, Its Nature and Scope," ACM SIGCAS Computers and Society 36, núm. 3 (2006): 21–36, https://doi.org/10.1145/1195716.1195719; Luciano Floridi, The Ethics of Information (Oxford: Oxford University Press, 2013), https://doi.org/10.1093/acprof:oso/9780199641321.001.0001.

9 Nithya Sambasivan et al., "'Everyone Wants to Do the Model Work, Not the Data Work': Data Cascades in High-Stakes AI," en Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Association for Computing Machinery, 2021), https://doi.org/10.1145/3411764.3445518.

10 Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," en Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (New York: Association for Computing Machinery,

para publicar, qué contenidos fueron preservados, qué plataformas son rastreables, qué idiomas predominan en ellas y qué criterios se aplicaron posteriormente para seleccionar información. La cantidad de datos no elimina estas características. En ocasiones puede ocultarlas bajo la apariencia de escala.

En este sentido, resulta necesario diferenciar entre disponibilidad y neutralidad ética. Un conjunto de datos puede estar disponible para descarga y, al mismo tiempo, contener errores de licencia, información sensible, poblaciones insuficientemente representadas o contenidos cuyo contexto dificulta ciertos usos. Las propuestas de documentación planteadas mediante *datasheets* y *data cards*[11] parten precisamente de que la existencia de un archivo no informa por sí misma sobre su motivación, composición, recopilación, procesamiento, usos recomendados y limitaciones.

La ética de la información de entrada puede entenderse entonces como el conjunto de principios y prácticas destinados a evaluar las condiciones bajo las cuales contenidos previamente existentes son convertidos en recursos para el entrenamiento de sistemas artificiales. Esta definición desplaza el análisis desde una pregunta limitada: ¿puede obtenerse el dato? hacia preguntas más amplias ¿de dónde procede?, ¿bajo qué condiciones fue producido?, ¿qué uso se autorizó?, ¿qué transformaciones ha sufrido?, ¿a quién representa?, ¿quién puede resultar afectado? y ¿quién debe responder por su incorporación?

No todas estas preguntas tendrán la misma respuesta jurídica en todas las jurisdicciones. Tampoco todos los problemas éticos son equivalentes a infracciones legales. Precisamente por ello resulta necesario mantener una distinción entre legalidad y legitimidad ética. Una práctica puede encontrarse en una zona jurídica todavía discutida y, aun así, requerir una evaluación sobre proporcionalidad, transparencia o respeto hacia quienes produjeron la información. Del mismo modo, una licencia técnicamente permisiva no garantiza que un conjunto sea representativo o de calidad.

El cambio fundamental consiste en abandonar una visión del dato como materia prima sin historia. En el contexto de la inteligencia artificial, los datos poseen genealogía. Reconstruirla no siempre será fácil, pero reconocer que existe es el primer paso para establecer responsabilidad.

III. Cuando la información se convierte en materia prima

La metáfora de la información como materia prima resulta útil y, al mismo tiempo, problemática. Es útil porque describe una característica de la economía digital contemporánea: contenidos producidos en múltiples contextos pueden ser recopilados a gran escala y transformados en recursos para sistemas computacionales. Es problemática porque puede inducir

2021), 610–23, https://doi.org/10.1145/3442188.3445922.

11 Las *datasheets for datasets* (fichas de documentación de conjuntos de datos) son documentos estandarizados que acompañan a un conjunto de datos y describen su motivación, composición, proceso de recopilación, tratamiento, usos previstos, limitaciones y mantenimiento, con el propósito de hacer visible su procedencia y las decisiones tomadas durante su construcción. Las *data cards* (tarjetas de datos) persiguen un objetivo similar, pero organizan esa documentación de manera estructurada y adaptable a diferentes tipos de usuarios, incorporando información sobre contexto, metodología, usos recomendados, riesgos, limitaciones y consideraciones éticas. Ambas propuestas buscan aumentar la transparencia y facilitar una evaluación responsable de los datos antes de utilizarlos en sistemas de inteligencia artificial. Véase Timnit Gebru et al., "Datasheets for Datasets," *Communications of the* ACM 64, núm. 12 (2021): 86–92, https://doi.org/10.1145/3458723; Mahima Pushkarna, Andrew Zaldivar y Oddur Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI," en *Proceedings of the* 2022 ACM *Conference on Fairness, Accountability, and Transparency*, 1776–1826 (New York: Association for Computing Machinery, 2022), https://doi.org/10.1145/3531146.3533231.

a pensar que esa materia se encuentra simplemente disponible, sin relaciones anteriores de autoría, privacidad, propiedad, atribución o contexto.

En los procesos de entrenamiento de gran escala, la recopilación automatizada permite reunir volúmenes que ningún equipo podría revisar documento por documento. El rastreo de la web, los repositorios públicos y la combinación de conjuntos preexistentes ofrecen una infraestructura eficiente para construir corpus. Sin embargo, la eficiencia de la recopilación no resuelve el estatuto de cada componente.

El estudio del corpus C4 permite ilustrar este punto. Al reconstruir su composición, Dodge y colaboradores encontraron contenidos de procedencias diversas y analizaron los efectos de los filtros empleados para conformarlo.[12] Lo relevante no es únicamente que el corpus contuviera información heterogénea. También demostraron que las decisiones adoptadas durante la limpieza podían producir exclusiones desiguales. Esto permite cuestionar la idea de que "limpiar" datos sea una operación puramente neutral.

Aquí aparece una distinción central para este artículo: accesibilidad no equivale a autorización.

La frase no pretende afirmar que todo contenido disponible públicamente requiera autorización individual para cualquier forma de análisis. Esa conclusión sería jurídicamente imprecisa y académicamente poco útil. Existen contenidos en dominio público, obras con licencias que permiten determinados usos, excepciones legales, información institucional destinada a reutilización y materiales expresamente publicados para favorecer minería, análisis o redistribución. Lo que la distinción señala es algo más básico: la posibilidad técnica de acceder a un contenido no permite inferir, por sí sola, las condiciones éticas y jurídicas de todos sus usos posteriores.

El problema se acentúa cuando los conjuntos de datos se derivan unos de otros. La investigación sobre procedencia y licenciamiento ha mostrado que las cadenas pueden incluir errores u omisiones que se propagan entre diferentes versiones y repositorios.[13] Si un conjunto es descargado, modificado, vuelto a publicar, combinado con otros y utilizado posteriormente para crear un corpus mayor, reconstruir las condiciones originales se vuelve progresivamente más difícil.

La cuestión del consentimiento introduce una dificultad adicional. En la web conviven contenidos publicados bajo muy diversas expectativas. Una persona puede escribir en un foro esperando interlocución pública sin imaginar que años después el texto formará parte de un sistema comercial. Una institución puede abrir un repositorio para facilitar consulta académica. Un fotógrafo puede permitir visualización pública mientras conserva determinadas restricciones de reutilización. Una biblioteca puede digitalizar patrimonio porque se encuentra en dominio público. Tratar estas situaciones como equivalentes simplemente porque pueden ser rastreadas por software elimina distinciones relevantes.

Los cambios recientes en las señales mediante las cuales los sitios web expresan restricciones relacionadas con la recopilación automatizada muestran que la infraestructura web no fue diseñada para expresar con claridad todas las preferencias que actualmente se esperan de ella.[14] Los términos de servicio, archivos destinados a robots y nuevas señales específicas para sistemas de IA pueden no coincidir. Esto

12 Dodge et al., "Documenting Large Webtext Corpora," 1286–1305.
13 Longpre et al., "Large-Scale Audit of Dataset Licensing," 975–87.
14 Longpre et al., "Consent in Crisis," 108042–87.

revela una crisis de mecanismos, no únicamente un conflicto entre creadores y desarrolladores. Entre ambos extremos se necesita gobernanza. Esto implica distinguir tipos de fuentes, condiciones de acceso, expectativas razonables, licencias, riesgos y finalidades. También implica reconocer que la escala importa. Un uso aislado y una recopilación masiva no tienen necesariamente la misma capacidad de afectar a individuos o comunidades.

La conversión de información en materia prima para inteligencia artificial debería entenderse, por ello, como un proceso documental sujeto a evaluación. Seleccionar datos equivale a seleccionar parte del mundo informacional que será incorporado al sistema. Excluir también constituye una decisión. El corpus final es el resultado de ambas operaciones.

IV. Los problemas éticos de alimentar una inteligencia artificial

Procedencia desconocida

La primera dificultad consiste en saber de dónde vienen los datos. En corpus construidos mediante agregación, la ruta entre el documento inicial y el conjunto final puede atravesar varios intermediarios. La procedencia puede perderse por ausencia de metadatos, por transformaciones sucesivas, por duplicación o por documentación insuficiente.

Las auditorías recientes sobre *datasets*[15] de inteligencia artificial muestran que este problema no es meramente hipotético.[16] La existencia de licencias ausentes, mal clasificadas o inconsistentes dificulta que desarrolladores posteriores determinen correctamente las condiciones de uso. Una cadena de datos sin buena documentación produce una cadena de incertidumbre.

Desde la ética de la información, esta pérdida tiene varias consecuencias. Dificulta reconocer a quienes produjeron contenidos, impide evaluar el contexto original, complica el retiro de información y limita la posibilidad de realizar auditorías posteriores. La procedencia debería entenderse por ello como atributo de calidad y como condición de responsabilidad.

Consentimiento y cambio de finalidad

El segundo problema es el cambio de finalidad. Una persona puede haber autorizado o aceptado una forma de publicación sin haber contemplado otras. Este punto no implica que toda reutilización requiera necesariamente consentimiento adicional; implica que la finalidad original constituye información éticamente relevante.

La diferencia puede observarse en el uso de datos personales. Que determinada información pueda ser encontrada públicamente no elimina de manera automática los riesgos derivados de su agregación. Datos dispersos pueden adquirir un significado diferente cuando se combinan. La recopilación a gran escala puede aumentar posibilidades de inferencia, identificación o reutilización.

15 Un *dataset* o conjunto de datos es una colección organizada de datos reunidos para su almacenamiento, análisis o procesamiento. En el ámbito de la inteligencia artificial, puede contener textos, imágenes, audios, registros numéricos u otros tipos de información que se emplean para entrenar, validar o evaluar modelos computacionales. Su composición, procedencia, calidad, representatividad y documentación influyen directamente en el comportamiento y las limitaciones de los sistemas desarrollados a partir de él.

16 Longpre et al., "Large-Scale Audit of Dataset Licensing," 975–87.

La UNESCO ha señalado expresamente la necesidad de proteger la privacidad y los derechos de las personas a lo largo del ciclo de vida de los sistemas de inteligencia artificial.[17] Este principio es importante porque rompe con la idea de que la ética aparece únicamente durante el uso final. Si la información personal entra al sistema bajo condiciones deficientes, el problema se origina antes del despliegue.

Propiedad intelectual y atribución

Los derechos de autor constituyen uno de los campos más visibles de la controversia contemporánea. Sin embargo, reducir la ética de los datos de entrenamiento a propiedad intelectual sería insuficiente. Una obra puede encontrarse en dominio público y seguir planteando problemas de representación o contexto. A la inversa, un contenido protegido puede estar sujeto a licencias, excepciones u otras condiciones que permiten determinados usos.

La cuestión ética central es más amplia: ¿qué obligaciones mantiene quien obtiene valor a partir de información producida por terceros?

La atribución proporciona un ejemplo. En la comunicación académica, reconocer las fuentes constituye una práctica central. Los modelos generativos complican este principio porque el entrenamiento opera sobre cantidades inmensas de información y no siempre permite relacionar una salida con documentos concretos. Esto no significa que la atribución sea técnicamente trasladable de forma directa desde un artículo académico al entrenamiento de un modelo. Sí significa que la pérdida de capacidad para identificar contribuciones no debería confundirse con inexistencia de esas contribuciones.

Calidad informativa

La tercera dimensión problemática suele recibir menos atención en debates centrados en derechos. Internet contiene conocimiento científico, información gubernamental, periodismo, obras culturales y materiales educativos, pero también errores, rumores, contenido manipulado, publicidad encubierta, duplicados, traducciones automáticas deficientes y páginas producidas para atraer tráfico.

Este problema adquiere una nueva capa con el crecimiento del contenido sintético. A medida que textos e imágenes generados por modelos se publican en la web, los futuros corpus pueden incluir cantidades crecientes de información producida por sistemas anteriores. La frontera entre contenido humano y sintético se vuelve menos evidente. La documentación de procedencia adquiere así una importancia adicional para poder conocer qué tipos de materiales integran nuevos ciclos de entrenamiento.

Sesgo y representación

Los datos web tampoco representan de forma proporcional a todas las poblaciones. Las investigaciones sobre documentación de datos y modelos lingüísticos han advertido que las tecnologías pueden producir resultados problemáticos cuando utilizan información procedente de unas poblaciones para sistemas que posteriormente serán aplicados a otras.[18]

Los estudios de grandes corpus también han mostrado que determinados mecanismos de filtrado pueden eliminar de manera desproporcionada textos

17 UNESCO, Recommendation on the Ethics of Artificial Intelligence.
18 Bender y Friedman, "Data Statements," 587–604; Bender et al., "Dangers of Stochastic Parrots," 610–23.

vinculados con grupos minoritarios.[19] En *datasets* multimodales, Abeba Birhane y colaboradores identificaron además contenidos explícitos y estereotipos dañinos que evidencian los riesgos de construir conjuntos masivos mediante procesos insuficientemente curados.[20]

El sesgo no consiste solamente en la presencia de contenido ofensivo. También puede expresarse mediante ausencia. Un modelo entrenado sobre una distribución lingüística desigual tendrá una base informacional diferente para cada lengua. Lo mismo ocurre con regiones, comunidades, tradiciones académicas o formas de conocimiento.

Esto permite formular una de las afirmaciones centrales del artículo: los modelos no sólo aprenden de información; aprenden dentro de las desigualdades de los sistemas de información de los que esa información procede.

La desigualdad informacional antecede a la inteligencia artificial. No todas las sociedades producen la misma cantidad de información digital, no todos los idiomas poseen la misma presencia en internet y no todas las instituciones cuentan con la misma capacidad de digitalización. Algunas comunidades aparecen abundantemente documentadas mientras otras dependen de fuentes externas que hablan sobre ellas.

Descontextualización

Existe también un problema menos visible: la pérdida de contexto. Una frase extraída de un foro, una noticia histórica o un texto satírico puede significar algo distinto cuando se separa del entorno donde apareció. Los procesos de entrenamiento necesitan transformar documentos en estructuras adecuadas para procesamiento, pero esa transformación puede disminuir señales contextuales.

La pérdida es especialmente relevante para información histórica. Un documento del siglo XIX que contiene expresiones discriminatorias posee valor como fuente para comprender una época. Su presencia en un corpus sin metadatos temporales o contextuales puede ser interpretada de otra manera por procesos posteriores. "Eliminar" todo contenido problemático tampoco es una solución simple, porque borraría evidencia histórica y cultural. El problema exige distinguir entre preservación, documentación y utilización.

Opacidad y trazabilidad limitada

La documentación se vuelve todavía más importante cuando los modelos son desarrollados por organizaciones que no publican suficiente información sobre sus datos. Las evaluaciones del *Foundation Model Transparency Index* han señalado niveles relevantes de opacidad en el ecosistema de modelos fundacionales, en particular respecto de los recursos utilizados para construirlos.[21]

La transparencia absoluta tampoco constituye una solución simple. Los conjuntos pueden contener información personal; publicar cada elemento podría agravar riesgos. Además, existen secretos comerciales y problemas de seguridad. La alternativa no es escoger entre revelar todo o no revelar nada. Es nece-

19 Dodge et al., "Documenting Large Webtext Corpora," 1286–1305.
20 Abeba Birhane, Vinay Uday Prabhu y Emmanuel Kahembwe, "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes," arXiv (2021), https://doi.org/10.48550/arXiv.2110.01963.
21 Rishi Bommasani et al., "The 2024 Foundation Model Transparency Index," Transactions on Machine Learning Research (2024); Alexander Wan et al., "The 2025 Foundation Model Transparency Index."

sario diseñar niveles de transparencia capaces de documentar las características relevantes sin reproducir daños.

Apropiación de valor

Finalmente, la inteligencia artificial generativa plantea una cuestión sobre distribución de valor. La construcción de sistemas capaces de generar productos comercialmente útiles depende, entre otros factores, de grandes cantidades de materiales producidos por comunidades humanas. Esa relación obliga a preguntar quién aporta valor y quién puede capturarlo.

No toda utilización de información exige compensación económica. Las bibliotecas, la investigación y el conocimiento abierto se sostienen precisamente mediante formas de circulación que no dependen de pago por cada consulta. Sin embargo, tampoco puede asumirse que todo contenido públicamente visible fue producido con la expectativa de contribuir a un servicio comercial.

La ética requiere reconocer la heterogeneidad de esas situaciones. Una fotografía con licencia comercial, un texto gubernamental de libre reutilización, una obra en dominio público, un artículo bajo determinada licencia abierta, una conversación personal y un libro protegido no deberían colapsarse dentro de una misma categoría denominada simplemente "datos disponibles".

V. La cadena de responsabilidad informacional

A partir de los problemas anteriores se propone una cadena de responsabilidad informacional para los datos de entrenamiento. El concepto parte de una premisa: la responsabilidad sobre un sistema generativo no comienza ni termina en quien interactúa con el modelo final. Existen decisiones previas relacionadas con la recopilación, organización, selección, licenciamiento, documentación y tratamiento de la información.

Procedencia → Legitimidad de uso → Calidad → Representación → Trazabilidad → Responsabilidad

El orden tiene una función analítica, aunque en la práctica las dimensiones pueden superponerse. No se pretende afirmar que cada dato atraviese linealmente seis departamentos o procesos técnicos. Se propone que cualquier estrategia de gobernanza de información para entrenamiento debería ser capaz de responder razonablemente a preguntas relacionadas con estas dimensiones.

1. Procedencia: ¿de dónde viene la información?

La primera dimensión es la procedencia. Consiste en identificar el origen de la información con el grado de precisión que resulte técnicamente posible y éticamente adecuado.

En un corpus pequeño podría documentarse cada elemento. En uno de gran escala puede resultar más realista documentar colecciones, dominios, repositorios, fechas o categorías. El nivel de granularidad puede variar, pero el principio permanece: no debería normalizarse que conjuntos fundamentales para desarrollar un sistema carezcan de información básica sobre su origen.

Esta dimensión recupera aportaciones previas sobre declaraciones de datos, fichas de *datasets*, tarje-

tas de datos y auditorías de procedencia.[22] Todas estas propuestas, desde perspectivas distintas, coinciden en que el contexto del dato importa.

La procedencia permite responder preguntas posteriores. Sin ella es difícil evaluar licencia, temporalidad, población representada, calidad o posibilidades de retiro. La procedencia funciona, por tanto, como una primera condición de auditabilidad.

Los metadatos mínimos podrían incluir, según el caso, nombre de la fuente o colección, responsable de su creación, método de obtención, fecha de recopilación, idioma, modalidad, transformaciones principales y relaciones con conjuntos anteriores.

2. Legitimidad de uso: ¿por qué puede utilizarse?

El segundo nivel consiste en determinar por qué su incorporación puede considerarse legítima.

La legitimidad incluye elementos jurídicos, pero no se agota en ellos. Deben considerarse derechos de autor, licencias, dominio público, términos aplicables, privacidad, expectativas razonables de las personas y finalidad del tratamiento. Esta dimensión es la que permite precisar la afirmación "accesibilidad no equivale a autorización". No significa que lo accesible nunca pueda utilizarse. Significa que la accesibilidad constituye una condición técnica, mientras que la legitimidad requiere información adicional.

Puede imaginarse una escala de situaciones. En un extremo se encuentran materiales expresamente ofrecidos para reutilización, dominio público o conjuntos creados específicamente para entrenamiento. En otro aparecen datos privados obtenidos sin autorización. Entre ambos existen numerosos casos: obras con distintas licencias, acceso abierto, páginas públicas con condiciones específicas, información de usuarios y contenidos sujetos a excepciones legales.

La responsabilidad consiste precisamente en no tratar toda esta diversidad como una única categoría.

El desarrollo de *Common Pile v0.1* resulta ilustrativo porque demuestra la posibilidad de construir a gran escala un corpus compuesto por materiales de dominio público y con licencias abiertas.[23] No elimina todos los problemas éticos, pero evidencia que la procedencia y las condiciones de uso pueden incorporarse deliberadamente al diseño del conjunto.

3. Calidad: ¿qué tan adecuada es la información?

La tercera dimensión pregunta si los datos son adecuados para el propósito. La calidad informativa comprende mucho más que ausencia de archivos corruptos. Incluye actualidad, exactitud, contexto, duplicación, presencia de información falsa, confiabilidad de las fuentes y efectos de los procedimientos de limpieza.

Un corpus puede ser jurídicamente utilizable y, al mismo tiempo, ser deficiente. Puede estar compuesto enteramente por materiales con licencias claras y contener sesgos extremos, errores o información obsoleta. Por ello, legitimidad y calidad deben evaluarse por separado.

También debe evitarse definir calidad exclusivamente como semejanza con un estándar lingüístico

22 Bender y Friedman, "Data Statements," 587–604; Gebru et al., "Datasheets for Datasets," 86–92; Pushkarna, Zaldivar y Kjartansson, "Data Cards"; Longpre et al., "Large-Scale Audit of Dataset Licensing," 975–87.

23 Nikhil Kandpal et al., "The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text," arXiv (2025), https://doi.org/10.48550/arXiv.2506.05209.

dominante. Los estudios sobre C4 muestran que ciertos filtros aparentemente asociados con la limpieza pueden tener consecuencias de representación.[24] La calidad no debería convertirse en un mecanismo que elimine variedades lingüísticas simplemente porque difieren de formas prestigiosas o frecuentes.

Resulta más adecuado hablar de adecuación. Un texto histórico puede ser excelente para estudiar el lenguaje de una época y problemático si se interpreta como descripción actual del mundo. Un foro informal puede ser inadecuado como fuente de hechos médicos y valioso para estudiar formas coloquiales de comunicación. La calidad depende también del uso.

4. Representación: ¿quién está y quién falta?

La cuarta dimensión se refiere a las distribuciones del corpus. La representación no exige que todos los conjuntos reproduzcan proporcionalmente a la población mundial. Tal objetivo sería imposible en muchos casos. Sí exige conocer, dentro de límites razonables, qué idiomas, regiones, tipos de fuentes y grupos tienen mayor o menor presencia.

Esta dimensión es importante porque las ausencias pueden ser tan significativas como las presencias. Una lengua con escasos recursos digitales puede quedar subrepresentada, aunque millones de personas la utilicen. Una comunidad predominantemente oral puede dejar pocas huellas disponibles para rastreo web. Un país con menor infraestructura de digitalización puede aparecer de manera fragmentaria.

Las desigualdades digitales se transforman así en desigualdades de entrenamiento. Por ello, la diversidad no debería evaluarse sólo mediante el número de idiomas declarado por un conjunto. Es necesario considerar volumen relativo, variedad de fuentes y calidad de la representación. Una lengua presente únicamente mediante traducciones automáticas no ocupa la misma posición informacional que otra representada mediante literatura, prensa, conversación, documentación institucional y producción científica.

La UNESCO ha señalado la necesidad de promover diversidad e inclusión en los conjuntos de entrenamiento y prestar particular atención a contextos caracterizados por escasez de datos.[25] La preocupación no es decorativa. Si la inteligencia artificial se convierte en infraestructura de acceso a la información, sus asimetrías pueden influir sobre quién recibe respuestas más precisas, qué referencias culturales reconoce y qué formas de conocimiento aparecen como centrales.

La evaluación de representación también debería preguntarse quién describe a quién. Una comunidad puede aparecer abundantemente en documentos producidos desde fuera y tener poca presencia mediante voces propias. Contar menciones no basta para determinar representación.

5. Trazabilidad y transparencia: ¿puede reconstruirse la trayectoria?

La quinta dimensión no equivale a procedencia. La procedencia identifica el origen; la trazabilidad permite seguir transformaciones. Un documento puede proceder de un repositorio concreto, pero después ser limpiado, fragmentado, mezclado, deduplicado,[26] reetiquetado y redistribuido. La trazabilidad busca registrar estas operaciones de manera suficiente para

24 Dodge et al., "Documenting Large Webtext Corpora," 1286–1305.
25 UNESCO, Recommendation on the Ethics of Artificial Intelligence.
26 La deduplicación es el proceso mediante el cual se identifican y eliminan registros, documentos o fragmentos repetidos

comprender cómo llegó desde una fuente hasta un conjunto determinado.

La importancia de esta dimensión aumenta cuando existe una cadena de intermediarios. Si un proveedor utiliza un conjunto creado por otro, necesita saber qué transformaciones previas ocurrieron. Si después combina ese material con otras fuentes, debería documentar las nuevas operaciones.

La transparencia asociada con esta dimensión debe ser significativa, no necesariamente total. Publicar terabytes de archivos sin explicación puede ser formalmente abierto y materialmente opaco. La transparencia requiere organización.

En este sentido, un documento de transparencia debería poder responder, al menos, qué tipos de fuentes fueron utilizados, de qué periodos proceden, cómo se obtuvieron, qué procedimientos de filtrado se aplicaron, qué materiales fueron excluidos, qué proporción aproximada corresponde a distintas modalidades o idiomas y qué limitaciones son conocidas.

La tendencia regulatoria reciente muestra una aproximación hacia esta idea. El *Reglamento de Inteligencia Artificial* de la Unión Europea establece obligaciones de documentación para proveedores de modelos de propósito general y exige un resumen suficientemente detallado del contenido utilizado para el entrenamiento.[27]

Puede denominarse a esto transparencia de la genealogía informacional. La expresión hace referencia a la posibilidad de explicar de manera estructurada de qué ecosistema de información surgió el sistema.

6. Responsabilidad y reconocimiento: ¿quién responde?

La última dimensión integra las anteriores y pregunta quién asume responsabilidad cuando aparecen problemas. Una dificultad frecuente en sistemas complejos consiste en dispersar responsabilidad entre numerosos actores hasta que ninguno parece ser responsable.

El creador del *dataset* puede afirmar que sólo recopiló materiales; el desarrollador, que utilizó un conjunto disponible; la plataforma, que sólo alojó archivos; el proveedor del modelo, que no puede conocer cada documento; y el usuario final, que únicamente consultó una herramienta.

La existencia de una cadena real de actores no debería convertirse en una cadena de exenciones. La responsabilidad debe asignarse de acuerdo con capacidad de decisión y control. Quien recopila información tiene obligaciones relacionadas con adquisición y documentación. Quien prepara un *dataset* debe registrar transformaciones y limitaciones. Quien integra múltiples conjuntos debe revisar las condiciones relevantes y conservar su trazabilidad. Quien desarrolla y comercializa un modelo debe realizar debida diligencia sobre las fuentes que utiliza. Quien despliega un sistema en un contexto particular debe evaluar riesgos propios de ese uso.

La responsabilidad es, por tanto, distribuida, pero no diluida. El reconocimiento forma parte de esta dimensión. No siempre será posible atribuir individualmente cada contribución informacional al comportamiento de un modelo, pero sí pueden existir

dentro de un conjunto de datos. En los sistemas de inteligencia artificial, esta práctica busca reducir redundancias, evitar la sobrerrepresentación de determinados contenidos y disminuir el riesgo de que el modelo memorice información duplicada durante el entrenamiento.

27 European Parliament and Council of the European Union, "Regulation (EU) 2024/1689."

mecanismos de reconocimiento a colecciones, comunidades, repositorios o fuentes. Cuando el desarrollo depende de conjuntos creados por otras instituciones, la documentación puede preservar ese vínculo.

También deben existir vías de rectificación. Una gobernanza responsable debería contemplar mecanismos para reportar información indebidamente incluida, solicitar revisión, corregir metadatos, actualizar licencias o eliminar futuras versiones cuando corresponda.

Las seis dimensiones forman una secuencia de preguntas: ¿De dónde viene? → ¿por qué puede utilizarse? → ¿es adecuada? → ¿a quién representa? → ¿**podemos reconstruir lo que ocurrió?** → ¿quién responde?

La utilidad de la cadena reside en que evita reducir todos los problemas a una sola categoría. Un conjunto puede superar una dimensión y fallar en otra. Puede tener licencia correcta, pero representación deficiente; buena calidad, pero procedencia insuficientemente documentada; diversidad amplia, pero información personal que no debió incorporarse.

VI. Información abierta no significa información libre de responsabilidad

El movimiento de acceso abierto transformó la comunicación científica al defender que la literatura académica pudiera circular con menos barreras. La *Budapest Open Access Initiative* formuló una definición amplia en la que el acceso abierto permite no sólo leer, sino también descargar, copiar, distribuir, buscar, enlazar y utilizar contenidos para otros fines lícitos, dentro de las condiciones correspondientes.[28]

El problema surge cuando se utiliza la palabra "abierto" sin atender a la licencia concreta. Existe una diferencia entre que un archivo pueda consultarse gratuitamente y que posea una licencia abierta determinada. También existen diferencias entre licencias abiertas. Algunas permiten usos comerciales, otras incorporan restricciones; algunas permiten obras derivadas bajo ciertas condiciones. Por ello, afirmar simplemente que un artículo "está en internet" proporciona muy poca información sobre su régimen de reutilización.

Esta distinción es particularmente importante para repositorios institucionales y bibliotecas digitales. Un repositorio puede ofrecer acceso a documentos cuyos autores conservan derechos. Puede albergar obras con distintas licencias e incluso materiales cuya consulta se permite bajo condiciones específicas. La infraestructura abierta no uniforma automáticamente el estatuto de todos sus objetos.

Lo mismo sucede con el dominio público. Una obra cuyo plazo de protección patrimonial ha expirado puede ser reutilizable desde el punto de vista de los derechos de autor, pero el conjunto que contiene su digitalización puede incorporar otros elementos. Además, el hecho de que un material sea reutilizable no resuelve cuestiones como calidad, sesgo o sensibilidad cultural.

Las bibliotecas y archivos ofrecen una analogía relevante. Preservar un documento no equivale a aprobar su contenido. Dar acceso a una fuente tampoco significa que todos sus posibles usos sean éticamente indistinguibles.

Las instituciones documentales han desarrollado durante décadas prácticas para contextualizar colecciones, proteger información sensible, describir procedencia y establecer restricciones justificadas.

28 Budapest Open Access Initiative, "Read the Budapest Open Access Initiative," Open Society Institute, 2002.

La inteligencia artificial podría beneficiarse de esa experiencia.

Una ética de los datos de entrenamiento debería distinguir, al menos, entre seis acciones: acceder, consultar, reproducir, reutilizar, realizar minería de textos y datos, y utilizar para entrenamiento.

Estas acciones pueden solaparse técnicamente, pero no son conceptualmente idénticas. La existencia de derechos para una no permite inferir automáticamente el estatuto de las demás en todas las jurisdicciones.

Esta diferenciación no tiene como propósito dar por terminado este análisis. Su finalidad es mejorar la seguridad jurídica y ética mediante una documentación más precisa. Un ecosistema donde las licencias y condiciones sean legibles para humanos y máquinas podría facilitar el desarrollo responsable más que un escenario de incertidumbre permanente.

El proyecto *Common Pile v0.1* proporciona un ejemplo de que es posible construir un corpus de gran escala con materiales de dominio público o expresamente licenciados de forma abierta.[29]

El proyecto no demuestra que toda inteligencia artificial pueda entrenarse exclusivamente con esa clase de materiales ni resuelve todas las controversias. Sí cuestiona la idea de que documentar licencias y procedencia sea necesariamente incompatible con construir corpus grandes.

Esta distinción también protege al propio movimiento de acceso abierto. Si "abierto" comienza a utilizarse como sinónimo de ausencia de cualquier responsabilidad, podrían aumentar los incentivos para cerrar contenidos por temor a usos no previstos. Las investigaciones sobre señales de consentimiento en la web han identificado precisamente un crecimiento de restricciones relacionadas con la inteligencia artificial.[30] La pérdida de confianza puede terminar reduciendo el bien informacional común.

VII. Hacia una ética de la inteligencia artificial desde la información

Durante los últimos años han surgido distintas propuestas para documentar *datasets*. Las declaraciones de datos, las fichas de conjuntos de datos y las tarjetas de datos comparten una preocupación por hacer visibles decisiones que de otro modo quedan ocultas.[31]

La cadena propuesta en este artículo articula estas prácticas desde la perspectiva de la ética de la información y plantea la necesidad de reconstruir la trayectoria de los conjuntos de datos, documentar las transformaciones que experimentan e identificar las responsabilidades asociadas a cada etapa de su integración y uso. De esta manera, la procedencia y la trazabilidad se incorporan como elementos centrales para una gestión responsable de la información utilizada en sistemas de inteligencia artificial.

En términos operativos, los conjuntos de entrenamiento deberían documentar, cuando corresponda y resulte técnicamente posible:

a) origen general y principales fuentes;

b) fecha o periodo de recopilación;

c) mecanismos de obtención;

d) licencias y condiciones identificadas;

29 Kandpal et al., "Common Pile v0.1."

30 Longpre et al., "Consent in Crisis," 108042–87.

31 Bender y Friedman, "Data Statements," 587–604; Gebru et al., "Datasheets for Datasets," 86–92; Pushkarna, Zaldivar y Kjartansson, "Data Cards."

e) modalidades de información incluidas;

f) idiomas y distribución aproximada;

g) criterios de inclusión y exclusión;

h) procedimientos de limpieza;

i) métodos de deduplicación;

j) presencia y tratamiento de información personal;

k) utilización de información producida por usuarios;

l) incorporación de datos sintéticos;

m) sesgos o limitaciones identificados;

n) transformaciones realizadas sobre datasets anteriores;

o) mecanismos para reportar errores de procedencia o licencia;

p) procedimientos para solicitar revisión o exclusión cuando corresponda.

La trazabilidad también debería mantenerse entre versiones. Los *datasets* cambian. Se incorporan nuevas fuentes, se eliminan documentos, se actualizan filtros y se corrigen errores. Una versión posterior puede ser sustancialmente diferente de la anterior. El registro de versiones permite saber qué corpus fue utilizado para qué modelo y en qué condiciones.

Esto adquiere especial importancia cuando una persona solicita la eliminación de información o cuando se identifica posteriormente un problema de licencia. Borrar un archivo de la versión actual no modifica automáticamente modelos ya entrenados. La transparencia debe dejar claro qué medidas son técnicamente posibles y cuáles no.

La responsabilidad informacional también necesita proporcionalidad ya que no todos los sistemas implican el mismo nivel de riesgo ni requieren idéntica documentación. Un modelo experimental entrenado con un pequeño corpus institucional controlado no plantea los mismos desafíos que un modelo de propósito general entrenado con grandes cantidades de contenido web. El reconocimiento institucional de estas necesidades está creciendo. La UNESCO establece que los Estados deberían promover estrategias de gobernanza que atiendan la calidad de los datos de entrenamiento y la adecuación de los procedimientos de recopilación y selección.[32] La Unión Europea avanzó posteriormente hacia obligaciones concretas de documentación para los modelos de propósito general.[33] Estos marcos no son idénticos a la cadena propuesta aquí, pero muestran una transición desde principios generales hacia obligaciones aplicadas a la información utilizada durante el entrenamiento.

Es importante, sin embargo, no confundir transparencia con solución completa. Documentar que un corpus contiene determinados problemas no vuelve automáticamente aceptable su uso. La transparencia permite identificar, evaluar y discutir. La legitimidad requiere decisiones adicionales. De igual forma, la regulación no sustituye la ética. Las leyes suelen establecer mínimos y operan dentro de jurisdicciones concretas. Los modelos y sus datos atraviesan fronteras. Una práctica puede cumplir formalmente una norma local y seguir produciendo impactos sobre comunidades ubicadas en otros contextos.

32 UNESCO, Recommendation on the Ethics of Artificial Intelligence.
33 European Parliament and Council of the European Union, "Regulation (EU) 2024/1689."

Aquí adquiere sentido la idea de una genealogía informacional de la inteligencia artificial. Cada sistema posee una historia formada por fuentes, decisiones, filtros, exclusiones y transformaciones. Conocer esa genealogía no implica reconstruir matemáticamente cada relación entre un documento y cada parámetro del modelo. Significa documentar la infraestructura social y documental que hizo posible su entrenamiento.

Esta perspectiva también modifica la noción de innovación. La velocidad de desarrollo no debería medirse únicamente por cuántos parámetros, datos o capacidades pueden añadirse. Un avance puede consistir en construir mejores mecanismos para identificar derechos, respetar preferencias, documentar fuentes, integrar lenguas poco representadas o producir *datasets* de mayor calidad.

Los profesionales de la información tienen un papel relevante en este proceso. Bibliotecólogos, archivistas, documentalistas y especialistas en gestión de información poseen experiencia en metadatos, procedencia, evaluación de fuentes, preservación, clasificación y derechos de acceso. La gobernanza de datasets no debería quedar restringida a ingeniería de software y derecho. Es también un problema de organización del conocimiento.

Incorporar esta perspectiva podría mejorar los equipos interdisciplinarios responsables de desarrollar inteligencia artificial. La pregunta técnica ¿cómo obtenemos más datos? se complementaría con preguntas informacionales: ¿qué datos necesitamos?, ¿qué representan?, ¿cómo fueron documentados?, ¿qué condiciones los acompañan? y ¿qué información estamos dejando fuera?

VIII. Conclusiones

La responsabilidad ética de la inteligencia artificial generativa (IAG) no comienza con la respuesta producida por el modelo, sino con las decisiones mediante las cuales información creada por personas, instituciones y comunidades se selecciona, recopila, transforma e incorpora a los sistemas. Desde esta premisa, el artículo propuso una cadena de responsabilidad informacional integrada por seis dimensiones: procedencia, legitimidad de uso, calidad, representación, trazabilidad y responsabilidad.

La propuesta permite evitar dos simplificaciones. La primera consiste en asumir que toda información técnicamente accesible puede emplearse para cualquier finalidad. La segunda supone que toda reutilización exige las mismas condiciones. Los conjuntos de entrenamiento reúnen materiales con regímenes, contextos y riesgos distintos, por lo que su evaluación requiere distinguir entre disponibilidad, legitimidad, calidad y adecuación.

Los problemas tampoco se reducen a los derechos de autor. Un corpus compuesto por materiales legalmente reutilizables puede mantener sesgos, excluir comunidades, incorporar información falsa o presentar deficiencias de documentación. Por ello, la transparencia debe ofrecer información significativa sobre fuentes, periodos, criterios de selección, transformaciones, idiomas, licencias, limitaciones y mecanismos de rectificación, sin asumir que documentar un problema basta para resolverlo.

La responsabilidad es distribuida entre quienes producen, recopilan, organizan, transforman y utilizan la información, pero no debe diluirse entre esos actores.

Su asignación debe corresponder a las decisiones y capacidades de control de cada participante en la cadena.

La ética de la información recuerda, finalmente, que los datos tienen historia. Antes de convertirse en material de entrenamiento fueron documentos, expresiones, registros o representaciones situadas en contextos humanos. Una IAG responsable requiere preguntar de qué información aprendió, cómo fue obtenida, qué condiciones acompañaron su uso, quién quedó representado o excluido y quién asumió la decisión de incorporarla.

La ética de la inteligencia artificial comienza, por tanto, antes de evaluar lo que el modelo produce: comienza en las condiciones bajo las cuales se decide de qué información puede aprender.

Referencias

Bender, Emily M., y Batya Friedman. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." Transactions of the Association for Computational Linguistics 6 (2018): 587–604. https://doi.org/10.1162/tacl_a_00041.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major y Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" En Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23. New York: Association for Computing Machinery, 2021. https://doi.org/10.1145/3442188.3445922.

Birhane, Abeba, Vinay Uday Prabhu y Emmanuel Kahembwe. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." arXiv, 2021. https://doi.org/10.48550/arXiv.2110.01963.

Bommasani, Rishi, Kevin Klyman, Sayash Kapoor, et al. "The 2024 Foundation Model Transparency Index." Transactions on Machine Learning Research, 2024. https://openreview.net/forum?id=38cwP8xVxD.

Budapest Open Access Initiative. "Read the Budapest Open Access Initiative." Open Society Institute, 2002. https://www.budapestopenaccessinitiative.org/read/.

Dodge, Jesse, Maarten Sap, Ana Marasović, et al. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." En Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 1286–1305. Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.98.

European Parliament and Council of the European Union. "Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)." Official Journal of the European Union, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Floridi, Luciano. "Information Ethics, Its Nature and Scope." ACM SIGCAS Computers and Society 36, núm. 3 (2006): 21–36. https://doi.org/10.1145/1195716.1195719.

Floridi, Luciano. The Ethics of Information. Oxford: Oxford University Press, 2013. https://doi.org/10.1093/acprof:oso/9780199641321.001.0001.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, et al. "Datasheets for Datasets." Communications of the ACM 64, núm. 12 (2021): 86–92. https://doi.org/10.1145/3458723.

Kandpal, Nikhil, Brian Lester, Colin Raffel, et al. "The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text." arXiv, 2025. https://doi.org/10.48550/arXiv.2506.05209.

Longpre, Shayne, Robert Mahari, Anthony Chen, et al. “A Large-Scale Audit of Dataset Licensing and Attribution in AI.” Nature Machine Intelligence 6 (2024): 975–987. https://doi.org/10.1038/s42256-024-00878-8.

Longpre, Shayne, Robert Mahari, Ariel Lee, et al. “Consent in Crisis: The Rapid Decline of the AI Data Commons.” Advances in Neural Information Processing Systems 37 (2024): 108042–108087. https://doi.org/10.52202/079017-3431.

Pushkarna, Mahima, Andrew Zaldivar y Oddur Kjartansson. “Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI.” En Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1776–1826. New York: Association for Computing Machinery, 2022. https://doi.org/10.1145/3531146.3533231.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen K. Paritosh y Lora M. Aroyo. “‘Everyone Wants to Do the Model Work, Not the Data Work’: Data Cascades in High-Stakes AI.” En Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, artículo 39, 1–15. New York: Association for Computing Machinery, 2021. https://doi.org/10.1145/3411764.3445518.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000380455

Wan, Alexander, Kevin Klyman, Sayash Kapoor, et al. “The 2025 Foundation Model Transparency Index.” arXiv, 11 de diciembre de 2025. https://doi.org/10.48550/arXiv.2512.10169.