

Chapter 6 / Capítulo 6



The Seven Deadly Sins of Artificial
Intelligence
ISBN: 978-9915-9928-6-0
DOI: 10.62486/978-9915-9928-6-0.Ch06
Pages: 130-154

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

Wrath: using artificial intelligence to cause harm

Ira: utilizar la inteligencia artificial para dañar

Libardo Ramón Polanía¹  

¹ Universidad de la Amazonia. Florencia, Colombia.

ABSTRACT

Technological wrath is used as a metaphor for studying how artificial intelligence can amplify harm, misinformation, violence, and discriminatory decisions. This chapter examines mechanisms of harm without reducing them to a single intention and preserves distinctions among objective, production, distribution, decision, and consequence. The analysis integrates studies of bias, toxicity, automated messaging, deepfakes, and public trust with institutional frameworks of responsibility. The findings show that a consequence may require correction even when attribution of harm remains uncertain, and that human oversight is meaningful only when reviewers have sufficient time, information, competence, and practical authority to change a decision. Provenance is distinguished from truthfulness: confirming where content came from can help reconstruct responsibility, but does not demonstrate that a claim is true. The chapter also examines harassment, automation of hate, identity manipulation, and erosion of trust, together with the need to respond to affected people rather than only to technical system performance. The synthesis connects prevention, support, and remedy with responsible actors, resources, and review procedures. An illustrative scenario involving a false academic instruction separates verification, care, and restoration of opportunities; its effectiveness has not been evaluated. The chapter concludes that responsibility is a verifiable relationship between evidence and the capacity to modify consequences. An adequate institutional response must identify mechanisms, reduce exposure, correct decisions, and repair harm without using uncertainty about intention as a reason for inaction. Care therefore requires technical correction to be accompanied by understandable procedures for the people who experience the consequences.

Keywords: artificial intelligence; algorithmic harm; disinformation; responsibility; remedy

RESUMEN

La ira tecnológica se utiliza como metáfora para estudiar cómo la inteligencia artificial puede amplificar daño, desinformación, violencia y decisiones discriminatorias. Este capítulo examina mecanismos de daño sin reducirlos a una intención única y conserva diferencias entre objetivo, producción, difusión, decisión y consecuencia. El análisis integra estudios de sesgo, toxicidad, mensajes automatizados, deepfakes y confianza pública, junto con marcos institucionales de responsabilidad. Los resultados muestran que una consecuencia puede requerir corrección aunque la atribución del daño permanezca incierta, y que la supervisión humana solo es significativa cuando quienes revisan disponen de tiempo, información, competencia y autoridad material para cambiar una decisión. Se distingue procedencia de veracidad: confirmar el origen de un contenido puede ayudar a reconstruir responsabilidad, pero no demuestra que la afirmación sea verdadera. El capítulo examina además acoso, automatización del odio, manipulación de identidades y erosión de la confianza, así como la necesidad de responder a personas afectadas y no solo al funcionamiento técnico del sistema. La síntesis relaciona prevención, atención y reparación con responsables, recursos y procedimientos de revisión. Una aplicación ilustrativa sobre una instrucción académica falsa separa comprobación, cuidado y restitución de oportunidades; su eficacia no ha sido evaluada. El capítulo concluye que responsabilidad significa una relación verificable entre evidencia y facultad de modificar consecuencias. Una respuesta institucional adecuada debe identificar mecanismos, reducir exposición, corregir decisiones y reparar daños sin convertir incertidumbre sobre intención en excusa para la inacción. El cuidado exige que la corrección técnica se acompañe de procedimientos comprensibles para quienes reciben las consecuencias.

Palabras clave: inteligencia artificial; daño algorítmico; desinformación; responsabilidad; reparación

INTRODUCCIÓN

La inteligencia artificial puede reducir el esfuerzo necesario para producir mensajes, adaptar representaciones y difundir contenidos. Estas capacidades también pueden utilizarse para hostigar, engañar o perjudicar. La metáfora de la ira estudia esa orientación del uso y las condiciones que amplifican el daño. El capítulo distingue acciones intencionales, errores de diseño y consecuencias distributivas. Sus mecanismos pueden relacionarse, aunque requieren responsabilidades y respuestas diferentes (Autio et al., 2024; Weidinger et al., 2021).

Un sesgo discriminatorio no implica necesariamente una intención de agredir; la ausencia de esa intención tampoco elimina el perjuicio. Una salida falsa puede resultar de un error generativo o de una operación deliberada de desinformación. Reconocer estas diferencias permite elegir controles pertinentes y evita atribuir emociones a un sistema. La evaluación debe examinar qué ocurrió, qué lo hizo posible, quién fue afectado y quién tenía capacidad de prevenir o corregir (Selbst et al., 2019).

La pregunta central es qué condiciones permiten identificar y corregir daños asociados a IA sin confundir intención, mecanismo y consecuencia. La respuesta debe considerar una cadena: definición de tarea, producción de salida, distribución, decisión y revisión. Una falla puede aparecer en cualquiera de estos momentos. El análisis utiliza la metáfora de la ira para estudiar capacidades e incentivos que amplifican perjuicios, sin atribuir emociones al modelo. Su precisión depende de describir lo ocurrido y conservar diferencias que permiten asignar una intervención pertinente (Raji et al., 2020; Suresh & Guttag, 2021; Weidinger et al., 2021).

La intención tiene un lugar y una exigencia de evidencia. Una salida puede perjudicar aunque nadie busque agredir, mientras una operación deliberada puede utilizar contenidos técnicamente correctos para hostigar o manipular contexto. La ausencia de intención no elimina responsabilidad por diseño y uso; la existencia de una falsedad tampoco identifica automáticamente a quien pretendió engañar. Estas distinciones permiten actuar sobre consecuencias comprobadas mientras se investiga atribución. La institución necesita procedimientos que protejan, revisen y comuniquen límites, con competencia y garantías pertinentes al caso (Selbst et al., 2019; Tabassi, 2023; UNESCO, 2021).

El campo educativo ofrece situaciones donde una salida adquiere autoridad sobre oportunidades. Una evaluación, una instrucción o un mensaje atribuido a una persona pueden modificar una trayectoria. La institución conserva responsabilidad por decidir qué función autoriza y cómo se revisa una discrepancia. La presencia de una herramienta puede añadir velocidad y escala, y la decisión sigue dependiendo de relaciones y recursos. El marco ético educativo permite atender a docentes, estudiantes y responsables de adopción, en lugar de limitar la evaluación a rendimiento técnico agregado (Holmes et al., 2022; Miao & Holmes, 2023; Zawacki-Richter et al., 2019).

La desinformación y la representación sintética requieren distinguir capacidad y operación real. Un experimento puede mostrar que ciertos mensajes convencen en una tarea y no demostrar cambios de conducta en una campaña. Un análisis prospectivo puede organizar riesgos y no medir ataques observados. El capítulo conserva diseños y estatutos para que una crítica no exceda evidencia. La respuesta incluye verificación de afirmaciones y examen de circulación, con responsabilidades distintas. Esta arquitectura permite reconocer utilidades de comunicación y atender mecanismos que convierten facilidad de generación en perjuicio (Autio et al., 2024; Goldstein et al., 2023; Spitale et al., 2023).

La confianza pública se examina como capacidad de atribuir y revisar con evidencia. La existencia de falsificaciones puede facilitar fabricación engañosa y negación de registros verdaderos. Ambas trayectorias afectan procedimientos. La institución debe conservar una ruta de comprobación y explicar qué se pudo establecer. El criterio de cuidado combina precisión y capacidad de respuesta: evitar exposición adicional, ofrecer apoyo y revisar consecuencias. La reparación no se deduce automáticamente de retirar una pieza o de reducir una métrica, y necesita seguimiento de efectos concretos (Chesney & Citron, 2019; Cortina, 2024; Raji et al., 2020).

La tesis propone que la responsabilidad sea una relación verificable entre información, autoridad y recursos. Un reporte debe poder modificar uso, registro o consecuencia. El capítulo ofrece categorías y una aplicación ilustrativa, sin resultados propios de implementación. La agenda posterior incluye estudiar accesibilidad de reclamación, distribución de errores y eficacia de correcciones. Las condiciones propuestas requieren evaluación contextual, con participación y privacidad. Su contribución consiste en hacer visible qué actor puede intervenir en cada etapa y qué evidencia necesita para que una declaración de cuidado produzca una acción examinable (Lachheb et al., 2023; Tabassi, 2023).

RESULTADOS Y DISCUSIÓN

TECNOLOGÍA AMPLIFICADORA Y CONDICIONES DEL CONFLICTO

CAPACIDAD, ESCALA Y RESPONSABILIDAD

Los conflictos tienen historias, actores e instituciones que exceden una aplicación. La IA puede modificar velocidad, alcance y costo de determinadas acciones. Ese cambio merece análisis sin convertirlo en explicación única de una violencia previa. La identificación de riesgos de lenguaje realizada por Weidinger y colaboradores diferencia discriminación, información, usos maliciosos y relación humano-computadora. Su clasificación ayuda a organizar preguntas sobre mecanismos, sin probar por sí sola la magnitud de cada efecto en una población (Weidinger et al., 2021).

La responsabilidad se distribuye a lo largo del proceso. Quien produce un contenido tiene obligaciones distintas de quien diseña una interfaz, organiza difusión o adopta una decisión. La distribución no debe diluir la posibilidad de actuar. El control institucional necesita identificar tareas, recursos y facultades de intervención. Si todas las partes atribuyen el problema a otra, la cadena puede mantener el daño aunque cada actor declare haber cumplido una función limitada (Raji et al., 2020).

EL DAÑO SIN INTENCIÓN MANIFIESTA

El estudio de Obermeyer y colaboradores mostró cómo una variable de costo utilizada para representar necesidad sanitaria podía producir una distribución desigual. El mecanismo importa porque permite revisar el objetivo y los datos, además de la precisión técnica. El caso no establece que todo algoritmo sanitario opere de la misma manera. Su contribución consiste en mostrar que una elección aparentemente operativa puede incorporar desigualdades sociales dentro del proceso de decisión (Obermeyer et al., 2019).

Los resultados de clasificación de género estudiados por Buolamwini y Gebru muestran otro mecanismo: diferencias de desempeño entre grupos en sistemas comerciales evaluados. Comparar ambos trabajos permite reconocer que la discriminación puede aparecer por representación, evaluación u objetivo. La respuesta necesita identificar el punto de origen y el contexto de uso. Exigir una única métrica de igualdad puede dejar fuera relaciones que producen el perjuicio (Buolamwini & Gebru, 2018; Selbst et al., 2019).

Suresh y Guttag ordenan fuentes de daño a lo largo del ciclo de aprendizaje automático, desde representación y medición hasta evaluación y despliegue. Ese marco permite examinar un resultado perjudicial sin atribuir una emoción al sistema ni presumir intención individual. La decisión sobre el problema y el contexto de uso puede crear un daño aunque el modelo funcione conforme a su objetivo técnico (Suresh & Guttag, 2021).

La ira como metáfora conserva aquí un alcance delimitado: usos e incentivos que permiten amplificar perjuicios. Los estudios de sesgo demuestran diferencias de error o asignación; no prueban por sí mismos una voluntad de atacar. El análisis distingue intención documentada, mecanismo de producción y consecuencia observable. Esa separación permite formular responsabilidades proporcionadas y evita transformar toda falla en una acusación indistinta.

EL MECANISMO NECESITA UNA HISTORIA DELIMITADA

Una explicación del daño debe reconstruir actividad, actores y decisiones. La capacidad técnica puede modificar un conflicto y no explicar su origen completo. Interesa identificar qué operación redujo esfuerzo, amplió alcance o convirtió una salida en consecuencia. Esa secuencia permite distinguir posibilidad general y mecanismo documentado. La taxonomía de Weidinger y colaboradores organiza riesgos, mientras Suresh y Guttag sitúan fuentes de daño durante el ciclo de aprendizaje automático. Ambos marcos aportan categorías para investigar; una afirmación sobre un incidente necesita evidencia del contexto y de lo ocurrido (Suresh & Guttag, 2021; Weidinger et al., 2021).

El resultado perjudicial puede originarse en objetivo, representación, medición, evaluación o despliegue. Estos niveles conservan relaciones distintas. Una salida exacta respecto de una variable puede utilizarse para una decisión que esa variable representa mal. El caso sanitario de Obermeyer y colaboradores ilustra un mecanismo asociado a costos como aproximación

de necesidad. La síntesis utiliza ese antecedente para preguntar por el objetivo y sus relaciones sociales, evitando trasladar el hallazgo a todos los algoritmos de salud o interpretar su existencia como intención individual de agredir (Obermeyer et al., 2019; Selbst et al., 2019).

LA MÉTRICA Y LA DECISIÓN PUEDEN SEPARARSE

Una medida técnica permite observar una propiedad bajo condiciones definidas. Su importancia depende de la decisión que recibe autoridad. El estudio de Buolamwini y Gebru documenta diferencias de clasificación entre grupos en sistemas comerciales evaluados; su tarea y fecha limitan generalización. Un procedimiento institucional debe estudiar cómo se utiliza esa clasificación y qué consecuencias produce. La precisión agregada puede ocultar distribución de errores, pero una comparación por grupos también necesita representación, muestras y criterios pertinentes. La evaluación requiere examinar ambos niveles antes de concluir equidad de una adopción (Buolamwini & Gebru, 2018; Selbst et al., 2019).

La documentación de conjuntos de datos y modelos permite reconstruir población, usos previstos y límites. Su valor depende de relación con la tarea institucional. Una tarjeta puede advertir que un sistema no fue evaluado para una población, y la organización debe decidir qué implica antes de usarlo. La revisión necesita capacidad para modificar configuración o suspender una función. Conservar documentos sin conectar hallazgos y decisión puede mejorar memoria y mantener el mismo riesgo. La auditoría organiza esa relación y requiere responsables con recursos efectivos (Gebru et al., 2021; Mitchell et al., 2019; Raji et al., 2020).

DEPENDENCIA Y SUPERVISIÓN MATERIAL

La presencia de una persona en el procedimiento no garantiza control suficiente. Puede carecer de tiempo, información o autoridad para revisar. Bainbridge examina paradojas de automatización; Parasuraman y Riley distinguen formas de uso; Lee y See relacionan confianza y dependencia apropiada. Estos antecedentes permiten preguntar qué puede hacer realmente quien supervisa. La organización necesita identificar cuándo revisa, con qué evidencia y qué ocurre si discrepa. La posibilidad nominal de intervención adquiere contenido cuando esas condiciones hacen practicable modificar una consecuencia (Bainbridge, 1983; Lee & See, 2004; Parasuraman & Riley, 1997).

La confianza también puede ser insuficiente y eliminar una asistencia útil. Dietvorst y colaboradores estudian aversión después de observar errores; Burton y colaboradores revisan decisiones aumentadas y factores que influyen en rechazo. Esta evidencia evita formular el cuidado como desconfianza obligatoria. La autorización debe ajustarse a función, desempeño y consecuencias, conservando apoyo pertinente. El capítulo propone comparar qué decisiones son adecuadas para asistencia y cuáles requieren otro procedimiento, con pruebas contextualizadas. Una falla documentada puede limitar una función y dejar abiertas otras posibilidades bajo condiciones diferentes (Burton et al., 2020; Dietvorst et al., 2015).

PARTICIPACIÓN ANTES DE FIJAR UNA CONSECUENCIA

Quienes realizan una actividad pueden identificar excepciones y efectos que una prueba general omite. El conocimiento tácito y explícito distinguido por Litvinenko permite examinar esa contribución. Los casos de voz laboral de Doellgast y colaboradores aportan un antecedente situado sobre intervención en decisiones. Su traslado requiere adaptación, pero sostiene una pregunta práctica: qué información reciben los actores y en qué momento pueden influir. La participación puede modificar objetivos o usos, además de reportar fallas después de una adopción ya fijada (Doellgast et al., 2026; Litvinenko, 2026).

La literatura organizativa sobre riesgos y uso de IA amplía dimensiones de seguridad, recursos y responsabilidades. Crocco y colaboradores revisan trabajos sobre riesgos de implementación; Fox y Solihin estudian procesos organizativos mediante entrevistas cualitativas. Sus diseños aportan mapas y experiencias, sin demostrar que una medida concreta reduzca causalmente cualquier daño. La síntesis utiliza estas fuentes para formular preguntas sobre capacidad de respuesta y coherencia de funciones. Una institución debe comprobar sus procedimientos y resultados antes de considerarlos mitigaciones efectivas (Crocco et al., 2026; Fox & Solihin, 2026).

El daño se vuelve investigable cuando se separan consecuencia, mecanismo e intención.

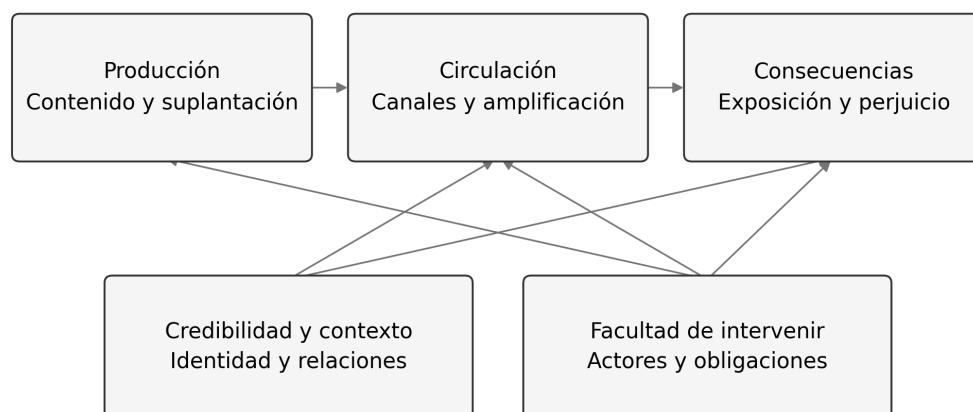
Puede existir evidencia suficiente para corregir una decisión mientras permanece incierto quién produjo una falsedad deliberadamente. La institución debe actuar sobre lo comprobado y conservar límites de atribución. Esta distinción permite cuidar sin improvisar acusaciones. Los marcos de riesgo y de auditoría orientan documentación y respuesta, mientras cada procedimiento necesita reglas y competencia pertinentes. La metáfora de la ira hace visible una capacidad de perjudicar y no sustituye el examen situado que permite identificar responsabilidades (Raji et al., 2020; Tabassi, 2023; UNESCO, 2021).

EL DAÑO SE ORGANIZA ENTRE PERSONAS Y HERRAMIENTAS

La metáfora de la ira describe una orientación de uso que busca intimidar, degradar o perjudicar, pero su análisis necesita reconstruir condiciones que permiten transformar esa intención en una actuación. Una herramienta puede reducir el esfuerzo de producir variantes, imitar identidades o adaptar mensajes; la circulación depende además de canales, incentivos y actores que seleccionan destinatarios. La capacidad de generar contenido y la capacidad de causar daño conservan una relación mediada por credibilidad, exposición y respuesta institucional. Esta distinción permite asignar obligaciones sin atribuir a un modelo una intención humana y sin reducir el problema a la voluntad del usuario final. Diseñadores, proveedores, intermediarios y organizaciones participan mediante decisiones diferentes que pueden facilitar o limitar una práctica. La responsabilidad se vuelve concreta cuando se identifican esas facultades y se pregunta quién podía prevenir una exposición, responder a una denuncia o corregir un contenido que seguía produciendo consecuencias (Chesney & Citron, 2019; Suresh & Gutttag, 2021; UNESCO, 2021).

La Figura 16 relaciona capacidades técnicas con condiciones de circulación y consecuencias. Producir una pieza resulta diferente de alcanzar a una audiencia y de perjudicar una relación; credibilidad y contexto ayudan a explicar esa trayectoria. Los actores con facultad de intervenir pueden modificar etapas distintas y adquieren obligaciones relacionadas con sus decisiones. El esquema permite discutir responsabilidad distribuida sin diluirla en una referencia genérica a la tecnología y sin asumir que detectar un origen artificial resuelve el perjuicio (Chesney & Citron, 2019; Suresh & Gutttag, 2021; UNESCO, 2021).

Figura 16. Relaciones entre producción, circulación y consecuencias del contenido dañino



Nota. Elaboración conceptual. Las flechas identifican relaciones argumentativas y facultades de intervención; no estiman probabilidades de daño ni atribuyen intención humana al modelo. Fuente: (Chesney & Citron, 2019; Suresh & Gutttag, 2021; UNESCO, 2021).

DESINFORMACIÓN AUTOMATIZADA

ERROR, ENGAÑO Y CIRCULACIÓN

Una afirmación falsa puede circular por desconocimiento o intención de engañar. La IA añade facilidad para producir variantes y adaptarlas a públicos. La respuesta debe reconocer esa capacidad y evitar deducir intención a partir de toda inexactitud. La verificación del contenido pregunta qué sostiene la afirmación; la investigación de una operación de desinformación necesita

examinar procedencia, coordinación y contexto. Ambos trabajos pueden complementarse, sin confundirse (Autio et al., 2024; Weidinger et al., 2021).

La fluidez aumenta plausibilidad y puede facilitar que una salida se reutilice sin revisión. TruthfulQA ofrece evidencia de modelos que reprodujeron concepciones falsas en una evaluación delimitada. El hallazgo permite entender un mecanismo de error que puede incorporarse a comunicación pública. Sus porcentajes históricos no deben trasladarse a cualquier sistema o contexto actual. La obligación editorial consiste en contrastar las afirmaciones relevantes antes de publicar, con controles proporcionados a las consecuencias (Lin et al., 2022).

Spitale, Biller-Andorno y Germani compararon mensajes breves sobre temas de información pública, escritos por personas o generados con GPT-3, y pidieron a participantes que evaluaran veracidad y procedencia. El estudio encontró ventajas de comprensibilidad para información verdadera generada y capacidad de producir desinformación convincente. La clasificación del origen también resultó difícil. El experimento demuestra efectos en esas tareas y textos; no prueba cambios de voto ni la eficacia de cualquier campaña desplegada (Spitale et al., 2023).

Goldstein y colaboradores organizan en un informe prospectivo las etapas de producción, acceso, difusión y formación de creencias. Las oportunidades de intervención cambian en cada etapa. Su análisis no constituye una medición de ataques observados, pero explica por qué mejorar un filtro de generación no resuelve necesariamente la distribución de contenidos ya producidos. La responsabilidad necesita seguir la cadena que permite que el material alcance a una población (Goldstein et al., 2023).

CORREGIR CON EVIDENCIA LOCALIZABLE

Una corrección útil identifica la afirmación, muestra su fundamento y explica qué cambia para el lector. Limitarse a etiquetar un contenido como falso puede dejar intacta la pregunta que motivó su circulación. La educación debe formar capacidad para localizar fuentes y reconstruir razones. La reflexión de Cortina sobre lenguaje y espacio público ofrece un criterio de deliberación: la confianza se sostiene cuando las personas pueden examinar y discutir lo que se presenta como verdadero (Cortina, 2024).

Las herramientas automáticas de análisis pueden asistir ese proceso y también introducir errores. La detección de confabulaciones mediante entropía semántica muestra posibilidades y límites del control técnico. Una salida consistente puede seguir siendo falsa, y una alerta requiere interpretación. El diseño debe combinar señales técnicas, fuentes y revisión pertinente, en lugar de presentar una clasificación como cierre definitivo de la discusión (Farquhar et al., 2024).

UNA FALSEDAD TIENE TRAYECTORIAS DIFERENTES

Un error puede generarse, copiarse, difundirse y recibir autoridad en momentos distintos. Investigar la cadena requiere identificar qué afirmación se presenta, qué fuente se invoca y qué acción solicita. La misma falsedad puede pasar de equivocación a uso deliberado si alguien conoce su condición y la emplea para engañar. Esta interpretación necesita evidencia, además del contenido. El capítulo distingue revisión editorial de investigación de coordinación o intención. La primera puede corregir una afirmación sin que la segunda esté resuelta, manteniendo una respuesta proporcionada al hallazgo (Goldstein et al., 2023; Weidinger et al., 2021).

La generación de lenguaje ofrece plausibilidad y puede reproducir concepciones incorrectas. TruthfulQA estudia ese problema en modelos y preguntas delimitados; Ji y colaboradores revisan alucinaciones; Maynez y colaboradores examinan fidelidad y factualidad de resumen. Las fuentes aportan mecanismos y condiciones diferentes. La evaluación debe comprobar qué salida se utilizó y qué evidencia faltó, evitando trasladar porcentajes históricos a servicios actuales. Una comunicación institucional necesita revisar afirmaciones relevantes, especialmente cuando su apariencia puede hacer que un destinatario actúe sin contraste (Ji et al., 2023; Lin et al., 2022; Maynez et al., 2020).

COMPENSIBILIDAD Y DAÑO POTENCIAL

Spitale y colaboradores encuentran que textos generados pueden comunicar información verdadera de manera comprensible y también producir desinformación convincente en su tarea experimental. Esta doble capacidad exige conservar beneficios y riesgos. La utilidad de una

explicación no determina verdad de todos sus contenidos, y el origen humano tampoco garantiza exactitud. La propuesta institucional utiliza criterios de evidencia aplicables a lo que se publica. El experimento conserva textos, participantes y actividad de clasificación; inferir eficacia de una campaña real requeriría estudiar difusión, exposición y conducta en otro diseño (Spitale et al., 2023).

La facilidad de adaptar un mensaje puede ampliar pertinencia o aprovechar expectativas de un público. El efecto depende de contenido, contexto y distribución. Goldstein y colaboradores ofrecen un análisis prospectivo de operaciones de influencia y posibles puntos de intervención. Su función es organizar preguntas, sin presentar una medición de ataques ocurridos. La síntesis propone identificar qué etapa permite actuar: preparación, acceso, publicación o revisión de creencias. Un filtro de generación puede aportar una parte y dejar intacta la circulación de materiales que ya existen (Autio et al., 2024; Goldstein et al., 2023).

SEÑALES TÉCNICAS Y FUENTES RECUPERABLES

La detección de incertidumbre por entropía semántica aporta una señal para ciertos errores en condiciones estudiadas. La consistencia de varias respuestas tampoco garantiza verdad. Farquhar y colaboradores ofrecen un antecedente para discutir posibilidades y límites, mientras la recuperación de una fuente permite examinar apoyo. Se propone una revisión que mantenga localizable la evidencia y explique la inferencia. Un aviso automático puede priorizar trabajo y necesita interpretación. La institución debe informar qué se comprobó antes de convertir una clasificación en una afirmación pública definitiva (Farquhar et al., 2024; Mitchell et al., 2019).

La confianza en una respuesta puede incorporar componentes cognitivos y emocionales. La encuesta de Q. Cheng y colaboradores diferencia esas dimensiones y ajuste a la tarea, sin establecer causalidad experimental general. Su aporte permite preguntar si cercanía o comodidad influyen en aceptación de un mensaje. La investigación de complacencia de M. Cheng y colaboradores añade consecuencias en conflictos interpersonales bajo versiones consultadas. La síntesis conserva ambos alcances y propone examinar cuándo satisfacción sustituye contraste, evitando atribuir a todo mensaje cordial un mismo efecto (M. Cheng et al., 2026; Q. Cheng et al., 2026).

CORREGIR LA AFIRMACIÓN Y SU CONSECUENCIA

Una rectificación útil identifica el contenido corregido, la fuente y lo que cambia para el destinatario. Puede requerir revisar una acción adoptada bajo el mensaje anterior. Una institución que publicó una fecha equivocada necesita aclararla y examinar oportunidades afectadas. Este ejemplo es una propuesta de análisis, sin incidente propio. La corrección debe circular en lugares pertinentes y conservar trazabilidad, con una intensidad ajustada al alcance conocido. Retirar el archivo puede reducir exposición y dejar pendientes consecuencias. La responsabilidad exige observar ambas dimensiones (Cortina, 2024; Raji et al., 2020; Tabassi, 2023).

La educación puede trabajar con fuentes y razonamiento, además de etiquetas. Long y Magerko y Ng y colaboradores aportan competencias de comprensión, evaluación y ética. Una actividad puede pedir que se identifique qué obra sostiene una afirmación y qué límite conserva. El resultado esperado es una decisión justificable, cuya eficacia debe evaluarse. La formación también necesita acceso a materiales y apoyo para interpretarlos. Las percepciones de estudiantes aportan información sobre expectativas y desafíos, mientras una tarea de verificación observa una capacidad diferente (Chan & Hu, 2023; Long & Magerko, 2020; Ng et al., 2021).

La respuesta a desinformación requiere conservar la posibilidad de reconocer hechos verdaderos. Una duda genérica sobre IA puede convertirse en motivo para descartar cualquier evidencia. El riesgo conceptual de negar registros auténticos señalado por Chesney y Citron muestra esta segunda trayectoria. La institución debe investigar la pieza y contexto, con fuentes y procedimientos pertinentes. La precisión permite corregir falsedades y evitar que la existencia de falsificaciones elimine toda capacidad de comprobación. La confianza pública necesita un recorrido revisable entre afirmación, evidencia y decisión (Autio et al., 2024; Chesney & Citron, 2019; Cortina, 2024).

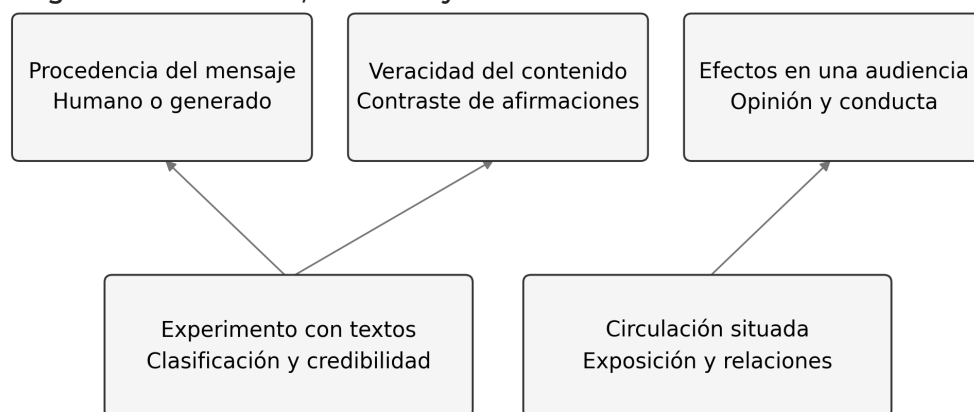
CREDIBILIDAD DEL MENSAJE Y EFECTOS DE UNA CAMPAÑA

La veracidad de un mensaje, la identificación de su origen y sus efectos sobre una audiencia

corresponden a niveles de análisis distintos. Un experimento con textos breves puede mostrar dificultades de clasificación y diferencias de credibilidad, mientras una campaña real introduce repetición, vínculos sociales y competencia con otras informaciones. La investigación de Spitale y colaboradores permite discutir resultados sobre mensajes humanos y generados sin afirmar que por sí sola explique cambios de opinión o de conducta. Para comprender desinformación automatizada interesa observar cómo una pieza se integra en secuencias de exposición y cómo aprovecha confianza previa en fuentes o identidades. La respuesta educativa puede fortalecer contraste de afirmaciones y atención al contexto, aunque su eficacia deba evaluarse en prácticas concretas. La distinción entre contenido y circulación evita atribuir toda influencia a la calidad lingüística y permite estudiar obligaciones de quienes amplifican materiales sin comprobar su procedencia o su sentido (Chesney & Citron, 2019; Long & Magerko, 2020; Spitale et al., 2023).

La Figura 17 distingue preguntas que requieren diseños y observaciones diferentes. Identificar quién produjo un mensaje atiende procedencia; contrastar lo que afirma examina veracidad; estudiar qué cambia en una audiencia requiere atender exposición y relaciones. Un resultado experimental sobre textos ayuda a comprender dificultades específicas y conserva límites al aplicarse a campañas. Esta distinción orienta una alfabetización que combina fuentes y contexto, manteniendo la necesidad de evaluar los efectos de intervenciones educativas en condiciones concretas (Chesney & Citron, 2019; Spitale et al., 2023).

Figura 17. Procedencia, veracidad y efectos como niveles de análisis diferentes



Nota. Elaboración conceptual. El experimento con mensajes breves aporta evidencias sobre clasificación y credibilidad. La influencia de una campaña o de una circulación real exige otra evidencia; el esquema no extrapola cambios de opinión o conducta. Fuente: (Chesney & Citron, 2019; Spitale et al., 2023).

VIOLENCIA DIGITAL, ACOSO Y AUTOMATIZACIÓN DEL ODIO

UNA CARGA CONCENTRADA EN LA PERSONA AFECTADA

La producción automatizada puede facilitar repetición y personalización de agresiones. El perjuicio depende de exposición, relaciones de poder y capacidad de respuesta, además del volumen de mensajes. La taxonomía de riesgos de lenguaje reconoce toxicidad y usos maliciosos como áreas distintas pero relacionadas. Su lectura permite considerar controles de generación y procedimientos de atención. El análisis debe describir el mecanismo sin reproducir instrucciones que favorezcan nuevas agresiones (Weidinger et al., 2021).

La respuesta centrada exclusivamente en bloquear cuentas puede exigir que la persona afectada asuma todo el trabajo de protección. Una institución necesita ofrecer un canal accesible, preservar evidencia necesaria y definir medidas de cuidado. El diseño debe limitar difusión adicional de información sensible. La privacidad sigue siendo una obligación durante la investigación: recopilar una denuncia no autoriza compartir indiscriminadamente sus detalles (Lachheb et al., 2023; UNESCO, 2021).

INCENTIVOS DE DIFUSIÓN

Las plataformas organizan visibilidad mediante sistemas y decisiones comerciales. La

relación entre recomendación y conflicto requiere evidencia específica; observar contenido polarizante no permite atribuir automáticamente su origen a un algoritmo. La evaluación debe preguntar qué objetivo se utiliza, qué señales se premian y qué controles existen. La perspectiva sociotécnica evita reducir una dinámica pública a una sola pieza técnica y permite incluir actores y procesos de decisión (Selbst et al., 2019).

La institución puede establecer criterios de publicación y moderación que protejan participación y admitan desacuerdo. Confundir crítica con agresión limita deliberación; tratar todo daño como ejercicio de expresión puede dejar desprotegidas a personas. La propuesta consiste en identificar conductas, consecuencias y procedimientos de revisión, con razones públicas y posibilidad de impugnar. El control debe mostrar cómo equilibra derechos en el contexto de la actuación (Cortina, 2024; UNESCO, 2021).

Gehman y colaboradores desarrollaron *RealToxicityPrompts*, con 100.000 inicios de texto derivados de la web, y examinaron continuaciones de modelos de lenguaje. Encontraron generación tóxica incluso desde entradas aparentemente inocuas y límites de las técnicas evaluadas para prevenirla. El resultado caracteriza modelos, datos y una medida de toxicidad de ese estudio. No identifica intención de los usuarios ni estima daños sufridos por una persona; la afectación concreta necesita evidencia adicional (Gehman et al., 2020).

El diseño de respuesta debe distinguir lo que observa una evaluación técnica de lo que relata una persona afectada. Un descenso de la métrica de toxicidad puede acompañar una mejora, pero no demuestra que se haya reparado la reputación de alguien ni que haya cesado el hostigamiento. La organización necesita un canal para atender esas consecuencias y revisar condiciones de distribución, además de medir las salidas de su modelo.

CONDUCTA, EXPOSICIÓN Y CONSECUENCIA

La violencia digital debe describirse mediante conductas y efectos. Repetición, representación no autorizada y difusión pueden relacionarse con daño, pero su análisis requiere contexto y evidencia. Una evaluación de generación tóxica mide una propiedad de salidas bajo un procedimiento; una denuncia describe una experiencia y necesita atención pertinente. *RealToxicityPrompts* aporta un antecedente técnico con datos y modelos delimitados. El capítulo conserva esa función y evita utilizar su métrica para identificar intención de una persona o estimar afectación individual (Gehman et al., 2020; Weidinger et al., 2021).

La carga puede concentrarse en quien recibe la agresión. Documentar, reportar y buscar ayuda exige tiempo y recursos, mientras el contenido puede seguir circulando. La institución debe organizar un canal comprensible y definir qué información necesita, sin multiplicar exposición. La ética del diseño de aprendizaje aporta una exigencia de privacidad que se mantiene durante respuesta a incidentes. La propuesta reconoce que cuidar incluye reducir trabajo innecesario de protección y hacer practicable acceso a apoyo, conservando procedimientos para examinar hechos y responsabilidades (Lachheb et al., 2023; UNESCO, 2021).

MODERACIÓN Y DELIBERACIÓN

La crítica, el desacuerdo y la agresión requieren distinciones que una puntuación aislada puede captar mal. El procedimiento debe identificar conductas y razones, con posibilidades de revisión. La perspectiva sociotécnica de Selbst y colaboradores permite examinar cómo una métrica cambia significado dentro de reglas y relaciones. La institución necesita observar consecuencias de moderación, incluida desigualdad de acceso a reclamar. Una reducción de contenido clasificado como tóxico puede representar una mejora parcial y seguir necesitando examen de participación y de errores de clasificación (Gehman et al., 2020; Selbst et al., 2019; Suresh & Gutttag, 2021).

La confianza en automatización puede aumentar aceptación de una señal sin revisión suficiente. Dzindolet y colaboradores estudian relaciones entre confianza y dependencia en tareas específicas; Lee y See aportan un marco de dependencia apropiada. La síntesis utiliza estos antecedentes para preguntar qué puede comprobar quien modera y cómo se registra una discrepancia. Una supervisión que requiere confirmar decisiones a gran velocidad puede limitar esa facultad. La responsabilidad necesita tiempo, información y autoridad, además de una persona formalmente asignada a la tarea (Bainbridge, 1983; Dzindolet et al., 2003; Lee & See, 2004).

PRIVACIDAD DE LA EVIDENCIA

Preservar un material pertinente difiere de difundirlo. El registro puede necesitar archivo, fecha y canal, con acceso limitado a quienes realizan una función. Recopilar conversaciones completas o información íntima adicional exige una finalidad y justificación que no se presumen desde la denuncia. Las fichas de datos ayudan a documentar composición y uso; el diseño ético exige minimizar tratamiento. La propuesta busca que la respuesta permita investigar y reduzca exposición, con decisiones explícitas sobre qué se conserva y quién puede acceder (Autio et al., 2024; Gebru et al., 2021; Lachheb et al., 2023).

La persona afectada debe conocer el recorrido del caso y qué decisiones están pendientes. La comunicación puede distinguir medidas de cuidado, hallazgos y atribuciones todavía inciertas. Esta claridad permite actuar sin prometer una conclusión que los datos no sostienen. La institución necesita evitar que una espera técnica deje sin apoyo o que una respuesta rápida introduzca una acusación insuficientemente examinada. Los marcos de gobernanza aportan orientación, mientras el procedimiento concreto debe adaptarse a sus obligaciones y capacidades. Su eficacia necesita observar consecuencias (Raji et al., 2020; Tabassi, 2023; UNESCO, 2021).

CONTEXTO EDUCATIVO Y BARRERAS DE ACCESO

Una ruta de atención puede ser nominalmente accesible y resultar difícil de utilizar por horarios, conectividad o lenguaje. La investigación regional de Kornhaber y colaboradores y las percepciones estudiantiles de Chan y Hu ofrecen preguntas sobre condiciones de participación, con poblaciones y objetos distintos. La institución debe estudiar sus propios canales y evitar atribuir los mismos resultados a sus usuarios. Se propone observar si las personas identifican dónde acudir y qué apoyo reciben, manteniendo esos indicadores separados de una tasa de incidentes o de una medida clínica (Chan & Hu, 2023; Holmes et al., 2022; Kornhaber et al., 2026).

La participación de trabajadores y estudiantes puede ayudar a reconocer cargas que una evaluación general omite. Los casos de voz laboral de Doellgast y colaboradores aportan un antecedente situado sobre intervención, mientras los estudios de autonomía y bienestar amplían preguntas organizativas. Su función aquí es orientar quién puede reportar un problema y modificar una práctica. El capítulo no traslada asociaciones de compras o enfermería a violencia digital. La propuesta exige información del procedimiento propio y una capacidad de respuesta que pueda comprobarse (Doellgast et al., 2026; Xie et al., 2026; Yue et al., 2026).

La respuesta será más defendible cuando una señal técnica, una experiencia y una consecuencia se relacionen sin confundirse. Se identifica qué información sustenta cada afirmación y qué acción permite. La prevención puede modificar diseño o circulación; la atención puede ofrecer apoyo; la reparación puede revisar una decisión o registro. Estos fines necesitan coordinación y recursos. La metáfora de la ira permite estudiar capacidad de amplificar perjuicio, mientras la evaluación conserva diferencias de intención, mecanismo y resultado que hacen posible intervenir con razones pertinentes (Floridi & Cows, 2019; Jobin et al., 2019; Mittelstadt, 2019; Raji et al., 2020).

LA EXPOSICIÓN DESIGUAL A LA VIOLENCIA DIGITAL

La violencia digital puede adquirir intensidad diferente según posición social, visibilidad y recursos para pedir apoyo. La facilidad para generar mensajes permite multiplicar piezas, pero la experiencia de daño depende también de su persistencia, de las relaciones del afectado y de la respuesta que encuentra. Una persona puede retirar una publicación y seguir afrontando circulación en otros espacios; otra puede contar con capacidad institucional para solicitar revisión rápida. Estas diferencias requieren atención a condiciones de desigualdad y evitan tratar la exposición como una experiencia uniforme. La posibilidad de explicar un caso debe conservar seguridad y privacidad, porque pedir evidencia repetida puede aumentar carga emocional o difusión de contenido dañino. El análisis de responsabilidad incluye preguntar qué información necesita realmente la institución, quién debe revisarla y cómo impedir que la búsqueda de comprobación reproduzca el perjuicio. La protección adquiere sentido cuando la respuesta modifica condiciones de exposición y conserva oportunidades para que la persona afectada recupere control (Fricker, 2007; Raji et al., 2020; Selbst et al., 2019; UNESCO, 2021).

DEEPFAKES COMO INSTRUMENTO DE DAÑO

REPRESENTACIÓN SIN AUTORIZACIÓN

Una representación sintética puede atribuir a una persona declaraciones o acciones que nunca realizó. Cuando involucra intimididad, el perjuicio puede persistir aunque su carácter falso se aclare. La ética de la respuesta debe reconocer dignidad y reducir exposición adicional. La autenticidad del archivo constituye una pregunta; el derecho de la persona a ser protegida frente a una representación dañina plantea otra. Ambas requieren atención (Autio et al., 2024; Weidinger et al., 2021).

La investigación debe preservar los materiales pertinentes en condiciones seguras y limitar su reproducción. Publicarlos nuevamente para demostrar falsedad puede ampliar el alcance del daño. Las instituciones necesitan asignar responsables de recepción, valoración y comunicación. El estudiante o trabajador afectado debe conocer el procedimiento y disponer de acompañamiento adecuado. La respuesta organizada evita que cada incidente dependa de improvisación y de la capacidad individual de insistir (Raji et al., 2020).

El caso de las llamadas de New Hampshire permite seguir una cadena documentada. El 21 de enero de 2024 circularon llamadas que imitaban mediante IA la voz de Joe Biden, entonces presidente de Estados Unidos, y desalentaban participar en la primaria próxima. Los documentos administrativos vinculan contenido sintético, identificación telefónica engañosa y transmisión. El problema excede la generación de una voz: su atribución y distribución pretendían conferir autoridad al mensaje (Federal Communications Commission [FCC], 2024a).

El acuerdo de agosto de 2024 con Lingo Telecom impuso una penalidad civil de un millón de dólares y un plan de cumplimiento que incluía controles de identidad de clientes y proveedores. En septiembre de 2024, la FCC anunció una sanción administrativa de seis millones de dólares contra Steve Kramer. Estas actuaciones documentan respuestas en dos puntos de la cadena. El análisis se limita a las actuaciones administrativas citadas de 2024; no presupone condena penal, estado posterior de cobro ni número de personas que cambiaron su conducta electoral (FCC, 2024a, 2024b).

La voz prestada corresponde al problema de atribución del acápite 4.4; este capítulo examina difusión, daño y respuesta. El informe estadounidense sobre réplicas digitales permite separar protección de identidad y protección de obras, mientras el caso muestra la necesidad de investigar intermediarios. La inferencia institucional consiste en verificar comunicaciones por canales independientes y asignar responsabilidades de respuesta; no en trasladar una sanción extranjera como regla automática para cualquier contexto (U.S. Copyright Office, 2024).

PROCEDENCIA Y COMUNICACIÓN RESPONSABLE

Verificar origen puede incluir contraste por canales independientes, examen técnico y revisión de contexto. Cada técnica tiene condiciones de aplicación y un alcance limitado. Una comunicación institucional debe explicar qué se pudo comprobar y qué permanece incierto. La duda sobre un archivo no autoriza atribuir intenciones a quienes aparecen ni desestimar automáticamente su situación. La precisión protege tanto confianza pública como garantías de las personas involucradas (Tabassi, 2023).

El aprendizaje ciudadano puede trabajar con materiales preparados y claramente identificados como ilustrativos. Comparar procedencias y afirmaciones permite comprender el problema sin exponer identidades reales. El propósito es formar criterio sobre evidencia y responsabilidad, no enseñar a perfeccionar suplantaciones. Esa orientación mantiene el foco pedagógico en el examen de contenidos y en los mecanismos de cuidado (Miao & Holmes, 2023).

La documentación de procedencia puede contribuir a reconstruir un incidente, pero la especificación C2PA no convierte una credencial en garantía de verdad. La comunicación institucional debe expresar qué fue comprobado: origen de un archivo, autorización de una persona, coincidencia con un registro o falsedad de una atribución. Un comunicado demasiado general puede crear una certeza que los datos no sostienen. La corrección necesita mantener localizable la evidencia que la respalda (Coalition for Content Provenance and Authenticity [C2PA], 2025).

LA FALSEDAD PUEDE PRODUCIR UNA CONSECUENCIA REAL

Una representación sintética puede ser falsa y afectar oportunidades, reputación o seguridad. La respuesta debe examinar lo que la pieza atribuye y lo que ocurrió después de su circulación. Aclarar origen puede corregir una creencia y dejar pendientes consecuencias. La ética de cuidado requiere observar exposición y acciones adoptadas, además de autenticidad del archivo. Los riesgos de lenguaje y de IA generativa aportan categorías amplias; investigar el incidente exige datos del contexto. El capítulo propone mantener esas dos escalas para evitar que una evaluación técnica cierre toda la respuesta (Autio et al., 2024; Weidinger et al., 2021).

La representación de identidad y la protección de una obra plantean preguntas diferentes. El informe estadounidense de 2024 sobre réplicas digitales ayuda a conservar esa distinción, con fecha y jurisdicción explícitas. Se utiliza como comparación conceptual, sin afirmar una norma universal o resolver derechos de un caso local. La institución puede establecer criterios éticos de autorización y presentación, mientras la determinación jurídica necesita procedimiento pertinente. Mantener esta frontera permite cuidar a una persona sin atribuir al documento citado una autoridad que su alcance no ofrece (U.S. Copyright Office, 2024; UNESCO, 2021).

LA CADENA DOCUMENTADA Y SUS LÍMITES

Las actuaciones administrativas de 2024 sobre llamadas en New Hampshire permiten relacionar una voz sintética con identificación y transmisión. El análisis conserva lo anunciado y documentado en esos actos, sin afirmar estado posterior de cobro, una conclusión penal o cambios electorales individuales. El caso aporta una cadena con puntos de responsabilidad distintos. Su transferencia a una institución educativa es una interpretación sobre verificación y respuesta, no aplicación automática de una sanción extranjera. La claridad sobre ese estatuto permite aprender de un antecedente sin exceder evidencia o jurisdicción (FCC, 2024a, 2024b).

Una comunicación puede utilizar identidad reconocible para adquirir autoridad. La institución necesita un recorrido donde el destinatario pueda confirmar instrucciones sensibles mediante un canal establecido. La propuesta debe especificar quién responde y qué ocurre durante la comprobación. Una opción nominal que nadie atiende ofrece poco apoyo. El diseño centrado en personas aporta criterios de control y explicación; las guías de interacción permiten revisar coherencia entre capacidad declarada y acciones disponibles. Afirmer eficacia de esa ruta requiere observar su funcionamiento en condiciones reales (Amershi et al., 2019; Shneiderman, 2020; Tabassi, 2023).

PROCEDENCIA COMO INSUMO REVISABLE

Los registros de C2PA pueden contribuir a examinar declaraciones de origen y cambios conforme a la versión técnica utilizada. Su interpretación debe conservar límites: ausencia de credencial no prueba falsedad y presencia no certifica verdad o consentimiento. La comunicación institucional identifica qué se pudo comprobar y qué permanece incierto. Una afirmación excesivamente amplia puede afectar a personas o procedimientos, incluso cuando pretende corregir. La responsabilidad requiere que la evidencia sea recuperable y que un hallazgo pueda revisarse cuando aparece información nueva (C2PA, 2025; Raji et al., 2020).

El examen técnico necesita competencia y condiciones de conservación. Se propone limitar reproducción del material y registrar decisiones sobre acceso, evitando ampliar exposición mientras se investiga. La privacidad permanece una obligación durante el proceso. El diseño ético del aprendizaje y las fichas de datos orientan finalidad y composición de registros, con funciones distintas. La institución debe explicar qué información solicita a la persona y por qué. El caso ilustrativo no introduce una técnica universal de detección ni promete que todas las piezas permitan atribución definitiva (Autio et al., 2024; Gebru et al., 2021; Lachheb et al., 2023).

COMUNICACIÓN Y RESTITUCIÓN DE OPORTUNIDADES

Una rectificación puede necesitar indicar que una instrucción carece de autorización o que una atribución fue comprobada como falsa. El alcance debe corresponder a hallazgos y canal. Si alguien perdió una oportunidad por actuar bajo el mensaje, la institución examina esa consecuencia mediante su procedimiento. Retirar un archivo no completa necesariamente la

respuesta. La propuesta vincula corrección informativa y revisión de decisiones, con responsables y recursos. El seguimiento pregunta si las medidas llegaron a destinatarios pertinentes y qué efectos siguen pendientes, sin inventar una reparación total (Cortina, 2024; Raji et al., 2020; UNESCO, 2021).

La formación puede utilizar materiales ilustrativos sin identidades reales para examinar afirmaciones y fuentes. El objetivo consiste en comprender qué evidencia se necesita, además de reconocer apariencia. La alfabetización incluye evaluación y ética, y debe observar decisiones del participante. Se propone un ejercicio de confirmar una instrucción y explicar qué sigue incierto. Su eficacia requiere evaluación propia. La actividad mantiene el foco en responsabilidad y cuidado, conservando alternativas de acceso y evitando que detectar señales perceptivas se convierta en una exigencia universal de competencia (Long & Magerko, 2020; Miao & Holmes, 2023; Ng et al., 2021).

La representación dañina exige una respuesta que siga archivo, atribución, distribución y consecuencia. Cada dimensión ofrece decisiones distintas y necesita evidencia propia. La síntesis propone una coordinación institucional que permita proteger, investigar y corregir, con límites de información. Los marcos de riesgos organizativos amplían preguntas sobre capacidad de implementación, sin garantizar mitigación. La metáfora de la ira identifica un uso que puede amplificar perjuicios; su evaluación permanece situada y requiere responsabilidad humana por los procedimientos que producen una consecuencia (Crocco et al., 2026; Fox & Solihin, 2026; Grupo de Trabajo de Frontera Tecnológica del Celam [GFT-CELAM], 2026; Tabassi, 2023).

UNA FALSIFICACIÓN PUEDE DAÑAR AUN DESPUÉS DE ACLARARSE

La aclaración de una falsificación puede corregir una afirmación y dejar abiertas consecuencias sobre reputación, seguridad o relaciones personales. Una representación sintética utiliza una identidad reconocible para introducir asociaciones que pueden permanecer en la memoria o reaparecer fuera del contexto de corrección. La evaluación del daño necesita observar esa trayectoria y reconocer que identificar un archivo como artificial constituye una operación limitada. También importa la autorización, porque una recreación correctamente etiquetada puede perjudicar a una persona cuyo consentimiento se omitió o excedió. La respuesta institucional debe distinguir medidas sobre contenido, sobre circulación y sobre apoyo al afectado. Las herramientas de procedencia aportan información cuando los registros se conservan; su pérdida o ausencia deja abierta la necesidad de contrastar otras evidencias. La discusión de deepfakes conduce así a una responsabilidad que continúa después de la detección y que requiere explicar qué consecuencias siguen activas y qué acciones pueden reducirlas (Chesney & Citron, 2019; C2PA, 2025; U.S. Copyright Office, 2024).

LA EROSIÓN DE LA CONFIANZA

DAÑO A PERSONAS Y A PROCEDIMIENTOS

El daño puede afectar una reputación y también un procedimiento de decisión. Si una organización recibe evidencias manipuladas y carece de examen, una falsedad puede convertirse en sanción. Si descarta todas las evidencias por sospecha, puede impedir protección. El problema exige un proceso que admita evaluación, contradicción y corrección. La confianza se sostiene mediante instituciones capaces de reconocer errores y explicar sus decisiones (Cortina, 2024).

La calidad de la información depende de criterios y relaciones. Una fuente identificable puede equivocarse; una corrección transparente puede fortalecer credibilidad. La revisión necesita conservar fecha, contexto y alcance, en lugar de valorar únicamente apariencia. El CELAM sitúa confianza entre los ecosistemas humanos afectados por la IA. Esa perspectiva permite reconocer que una tecnología participa en vínculos cuyo deterioro tiene consecuencias más amplias que una respuesta incorrecta (GFT-CELAM, 2026).

Chesney y Citron advierten que la existencia de falsificaciones también puede servir para negar pruebas auténticas. La institución debe conservar la posibilidad de confirmar hechos verdaderos, junto con la capacidad de retirar una atribución falsa. El daño afecta tanto a una persona como al procedimiento que permite reconocer evidencia. El análisis conceptual fundamenta esa doble obligación sin establecer una tasa de afectación para todos los contextos (Chesney & Citron, 2019).

DECISIONES DE ALTO IMPACTO

En áreas que afectan salud, seguridad o derechos, la evaluación debe ser más exigente. La OMS sitúa gobernanza y responsabilidad en aplicaciones sanitarias. El ICRC plantea restricciones y control humano sobre sistemas de armas autónomos, un ámbito particular que conserva obligaciones jurídicas y éticas específicas. La comparación muestra que el grado de automatización debe examinarse respecto de consecuencias y posibilidad de intervención; no existe una respuesta idéntica para todas las tareas (International Committee of the Red Cross [ICRC], 2021; World Health Organization [WHO], 2021).

El criterio de cuidado exige reconocer decisiones que una institución no debe automatizar en las condiciones disponibles. Esa conclusión puede ser temporal y revisable. Un límite razonado no implica rechazar toda asistencia; permite identificar qué información, competencias o salvaguardas faltan. Documentar la decisión ayuda a distinguir prudencia de mera inercia y conserva la posibilidad de reconsiderar cuando existen mejores condiciones (Tabassi, 2023).

CONFIAR EN UN PROCEDIMIENTO QUE PUEDE CORREGIRSE

La confianza institucional se sostiene cuando existe una relación comprensible entre evidencia y decisión. Una persona necesita conocer qué se afirma, qué datos lo respaldan y cómo puede solicitar revisión. La proliferación de contenidos sintéticos añade dificultades y no elimina esa exigencia. Cortina sitúa la deliberación en razones que pueden discutirse; el marco de auditoría conecta documentación con responsabilidad. La síntesis propone una confianza revisable, donde reconocer un error y actuar sobre sus consecuencias forma parte de la credibilidad del procedimiento (Cortina, 2024; Raji et al., 2020).

El daño puede aparecer por aceptar una falsedad o por descartar una prueba verdadera. El riesgo conceptual de negar registros auténticos descrito por Chesney y Citron permite considerar ambas trayectorias. La organización debe examinar la pieza y contexto, en vez de usar posibilidad de IA como motivo suficiente para cerrar una investigación. Conservar evidencia y explicación ayuda a mantener capacidad de comprobación. La conclusión debe indicar qué pregunta se respondió: origen, contenido, autorización o consecuencia. Esa precisión evita que una certeza parcial se extienda a todo el caso (Autio et al., 2024; Chesney & Citron, 2019; C2PA, 2025).

DESIGUALDAD DE RECURSOS PARA RECLAMAR

Una ruta de revisión puede exigir tiempo, conexión, lenguaje técnico o documentos que algunas personas no pueden obtener. La igualdad formal de procedimiento debe contrastarse con capacidad de uso. Los antecedentes de brechas digitales en Uruguay y de experiencias regionales australianas aportan preguntas, con objetos y contextos propios. La institución necesita estudiar sus participantes y ofrecer apoyos pertinentes. La propuesta incluye observar quién consigue presentar una explicación y quién abandona por obstáculos, manteniendo esa información separada de un juicio sobre veracidad del reclamo (González-Chouciño & Ruiz-Callado, 2023; Kornhaber et al., 2026; UNESCO, 2021).

La evaluación de escritura de Cui y las investigaciones de detectores de Liang y Weber-Wulff y colaboradores permiten considerar otro riesgo: una señal automatizada puede convertirse en consecuencia académica con errores distribuidos de manera desigual. Los materiales, herramientas y fechas limitan inferencia a versiones actuales. La síntesis utiliza esos antecedentes para discutir autoridad de evaluación y oportunidad de respuesta. Un porcentaje necesita relación con pregunta y evidencia, y la institución conserva responsabilidad por revisar tanto la entrega como el procedimiento que produjo sospecha (Cui, 2026; Liang et al., 2023; Weber-Wulff et al., 2023).

ALTO IMPACTO Y FUNCIÓN DE LA ASISTENCIA

Las consecuencias determinan qué información y competencia se necesitan antes de autorizar una función. La OMS propone gobernanza en salud; el ICRC formula una posición de 2021 sobre armas autónomas. Sus ámbitos y obligaciones difieren y se mantienen explícitos. La comparación permite reconocer que algunas decisiones requieren controles y límites específicos. El capítulo no ofrece una autorización universal de automatización ni un juicio jurídico sobre

tecnologías actuales. Su argumento examina qué capacidad de intervenir permanece cuando una salida se transforma en una consecuencia importante (ICRC, 2021; Tabassi, 2023; WHO, 2021).

La supervisión exige observar el trabajo real. Bainbridge y los marcos de automatización muestran que una persona puede quedar encargada de situaciones difíciles con información o práctica insuficientes. Una decisión institucional debe precisar cómo mantiene competencia y qué apoyo recibe. La revisión humana puede aportar interpretación situada y también cometer errores; su presencia no constituye una garantía absoluta. La propuesta consiste en evaluar el procedimiento completo, con tareas, recursos y facultades, conservando la posibilidad de limitar una función cuando esas condiciones resultan insuficientes (Bainbridge, 1983; Lee & See, 2004; Parasuraman & Riley, 1997).

BIENESTAR Y CAPACIDAD DE PARTICIPAR

La confianza depende también de relaciones laborales y educativas. Si reportar un problema se interpreta como resistencia injustificada, pueden ocultarse dificultades de uso. Los casos de voz laboral de Doellgast y colaboradores ofrecen un antecedente situado sobre participación, mientras los estudios de bienestar y autonomía amplían preguntas sobre roles. La síntesis no traslada sus resultados a cualquier institución. Se propone identificar qué efecto puede tener una observación y cómo se comunica una corrección. Una cultura de revisión requiere recursos y una oportunidad practicable de expresar desacuerdo (Doellgast et al., 2026; Xie et al., 2026; Yue et al., 2026).

Las dimensiones emocionales, sociales y cognitivas del bienestar descritas en el marco de Filippelli y colaboradores permiten ampliar qué consecuencias se preguntan. El acceso limitado a extractos no sostiene magnitudes ni efectos causales de aplicación. Su papel consiste en evitar que una medida técnica cierre toda evaluación de experiencia. La institución debe producir evidencia de su contexto y observar si una respuesta reduce exposición, restituye una oportunidad o crea carga adicional. Los indicadores deben conservar función y no transformarse en diagnóstico individual desde una señal aislada (Filippelli et al., 2026; Holmes et al., 2022; Lachheb et al., 2023).

La confianza defendible combina capacidad de confirmar, retirar y reparar. Estas acciones requieren información y autoridad asignada. La propuesta puede fallar si el procedimiento es incomprensible, si los datos son insuficientes o si quienes revisan carecen de recursos. Esas condiciones forman parte del análisis y de la agenda de evaluación. Los principios éticos ofrecen orientación, pero deben convertirse en decisiones observables. La institución necesita comprobar qué ocurre cuando alguien cuestiona una salida y si la respuesta modifica efectivamente la consecuencia que produjo (Floridi & Cows, 2019; Jobin et al., 2019; Mittelstadt, 2019; Shneiderman, 2020).

LA DESCONFIANZA GENERAL TAMBIÉN TIENE COSTOS

La posibilidad de falsificar contenidos puede utilizarse para desacreditar documentos auténticos y erosionar condiciones compartidas de comprobación. Ese efecto amplía el problema más allá de piezas concretas: si cualquier registro se descarta por la mera existencia de herramientas generativas, la institución pierde recursos para reconocer hechos y responder a reclamaciones. La confianza necesita sostenerse mediante procedimientos comprensibles que permitan contrastar sin exigir certeza imposible. Un registro de procedencia aporta una dimensión, mientras fuentes independientes, contexto y posibilidad de réplica ayudan a interpretar su significado. La intensidad de la comprobación debe ajustarse a consecuencias y a recursos disponibles, porque una exigencia desproporcionada puede excluir a quien tiene menos capacidad para documentar su experiencia. El objetivo consiste en conservar una relación entre evidencia, reconocimiento y corrección. Una política que solo aumenta sospecha puede facilitar el mismo deterioro que busca evitar si debilita las condiciones para escuchar y evaluar afirmaciones pertinentes (Chesney & Citron, 2019; C2PA, 2025; Fricker, 2007).

RESPONSABILIDAD HUMANA FRENTE AL DAÑO ALGORÍTMICO

PREVENCIÓN, RESPUESTA Y REPARACIÓN

La responsabilidad incluye revisar riesgos antes del despliegue, observar resultados durante

el uso y responder cuando ocurre un daño. Cada etapa necesita personas con facultades claras. La auditoría de Raji y colaboradores organiza documentación a lo largo del ciclo institucional. Su utilidad depende de que los hallazgos puedan modificar decisiones. Un control que registra problemas y carece de capacidad de intervención deja abierta la brecha que pretende cerrar (Raji et al., 2020).

La reparación requiere examinar consecuencias para la persona, corregir registros y evitar repetición. También necesita reconocer que retirar un contenido no elimina siempre los efectos. El diseño institucional debe contemplar comunicación, apoyo y revisión del procedimiento que permitió el daño. La propuesta de esta obra vincula cuidado con acciones verificables: quién recibe el caso, qué puede decidir y cómo se comprueba que la respuesta alcanzó su objetivo (Autio et al., 2024; UNESCO, 2021).

APLICACIÓN ILUSTRATIVA ANTE UNA SUPLANTACIÓN

Una institución educativa recibe una denuncia por una grabación que atribuye declaraciones ofensivas a un estudiante. En el escenario propuesto, preserva el archivo en un entorno restringido, recoge información necesaria y evita difundirlo. Un equipo competente examina procedencia y contexto. Mientras se esclarece el caso, se adoptan medidas de cuidado proporcionales y se conserva un procedimiento que permite a las personas involucradas aportar información y controvertir interpretaciones.

La comunicación final distingue lo comprobado de las hipótesis y explica las correcciones. La institución revisa cómo circuló el material y qué controles fallaron, sin convertir esa revisión en vigilancia general. El seguimiento pregunta si cesó la exposición y si las medidas atendieron las necesidades de la persona afectada. El escenario traduce responsabilidad en una secuencia de acciones, con límites y criterios de revisión, en lugar de confiar toda la respuesta a una declaración sobre uso ético (Lachheb et al., 2023; Raji et al., 2020).

En un escenario educativo, la circulación de un audio atribuido a una autoridad que modifica una fecha de evaluación requiere una respuesta organizada. La persona responsable confirma la instrucción por un canal institucional independiente, conserva el archivo y registra dónde se difundió. El equipo de comunicación publica una rectificación precisa en esos mismos espacios y ofrece a quienes actuaron bajo el mensaje una vía para revisar consecuencias. Son decisiones ilustrativas; no se presenta un incidente real de una universidad.

La respuesta distribuye trabajo: quien conserva evidencia limita el acceso al material, quien verifica separa hechos de hipótesis y quien decide reparación examina las consecuencias concretas. El registro incluye alcance conocido de difusión, momento de rectificación y medidas pendientes. Cuando la identidad no puede establecerse todavía, puede suspenderse la aplicación del mensaje sin acusar a una persona. Corregir significa restituir oportunidades afectadas y modificar el procedimiento que hizo posible el daño; retirar el archivo no completa necesariamente esa tarea.

RESPONSABILIDAD QUE CONSERVA UN PUNTO DE DECISIÓN

Distribuir funciones puede mejorar respuesta y también diluir autoridad. La institución necesita identificar quién recibe información, quién interpreta y quién decide una corrección. Cada responsable debe conocer límites y recursos. La auditoría de Raji y colaboradores organiza procesos durante el ciclo y requiere que los hallazgos puedan modificar uso. NIST articula gobernanza, mapeo, medición y gestión. La síntesis utiliza estos marcos para conectar registro con acción, evitando considerar suficiente una cadena donde todos documentan y nadie puede cambiar la consecuencia (Raji et al., 2020; Tabassi, 2023).

Una respuesta puede iniciar cuidado mientras se verifica atribución. Debe declarar qué medida responde a exposición conocida y qué conclusión sigue pendiente. Esta distinción permite actuar con prudencia y conservar garantías de las personas involucradas. La institución necesita reglas aplicables y competencia pertinente, además de una secuencia técnica. La propuesta del capítulo define relaciones entre prevención, atención y reparación, sin imponer un procedimiento jurídico universal. Su eficacia debe observarse en resultados y cargas de quienes participan (Cortina, 2024; Lachheb et al., 2023; UNESCO, 2021).

PREVENCIÓN VINCULADA CON LA TAREA

La prevención comienza por delimitar función y consecuencias. Un asistente puede ayudar a organizar información y requerir otro procedimiento para decidir una sanción o un apoyo. La documentación de modelos y datos permite revisar usos previstos y poblaciones; la institución añade pruebas de su tarea. La revisión educativa de Zawacki-Richter y colaboradores y el marco de Holmes y colaboradores apoyan participación de actores del contexto. La configuración debe examinarse por resultados y distribución, evitando autorizar decisiones solo porque una capacidad general funciona en una demostración (Gebru et al., 2021; Holmes et al., 2022; Mitchell et al., 2019; Zawacki-Richter et al., 2019).

La formación puede utilizar ejemplos donde una afirmación cambia de significado según evidencia. Se propone distinguir error, intención y consecuencia y justificar qué acción corresponde. Long y Magerko y Ng y colaboradores aportan competencias de evaluación; las guías universitarias revisadas por Ares-Ferreirós y colaboradores aportan recomendaciones de uso. Sus contribuciones no prueban eficacia de la actividad propuesta. La evaluación debe observar si los participantes identifican límites y utilizan un recorrido de revisión, además de recordar definiciones o declarar satisfacción con el curso (Ares-Ferreirós et al., 2026; Long & Magerko, 2020; Ng et al., 2021).

ATENCIÓN Y REPARACIÓN CON OBJETIVOS DISTINTOS

La atención puede reducir exposición y ofrecer apoyo. La reparación puede corregir un registro, revisar una decisión y modificar el proceso que permitió el daño. Retirar contenido puede contribuir y dejar consecuencias pendientes. La institución necesita acordar qué resultado busca cada medida y cómo lo observará. La propuesta conserva participación de la persona afectada y evita exigir información íntima innecesaria. Los marcos éticos orientan dignidad y autonomía, mientras el seguimiento debe comprobar si los recursos ofrecidos resultan utilizables y pertinentes en el contexto (Floridi & Cowls, 2019; Lachheb et al., 2023; UNESCO, 2021).

La revisión del procedimiento puede identificar datos inadecuados, una variable que representa mal la necesidad o una autoridad de salida excesiva. Los estudios de sesgo muestran mecanismos distintos que requieren controles propios. La corrección debe relacionarse con origen del problema, evitando una medida genérica que deja intacta la decisión. Se propone registrar la hipótesis de falla, la evidencia y el cambio realizado, y observar si modifica resultados. La persistencia de un daño puede exigir revisar el objetivo completo, incluso si mejora una métrica técnica (Buolamwini & Gebru, 2018; Obermeyer et al., 2019; Selbst et al., 2019; Suresh & Guttag, 2021).

INFORMACIÓN SUFICIENTE Y TRANSPARENCIA PROPORCIONADA

La comunicación debe explicar hechos comprobados, límites y acciones, con cuidado de identidades y datos. Una explicación técnica extensa puede omitir lo que la persona necesita saber para actuar. Las guías de interacción de Amershi y colaboradores permiten examinar claridad y posibilidad de corregir. La propuesta institucional identifica qué decisión puede revisar el destinatario y qué canal existe. Su acceso debe comprobarse, porque una fórmula de contacto sin respuesta puede aumentar carga. La transparencia adquiere valor cuando permite comprender y cambiar, además de publicar información (Amershi et al., 2019; Raji et al., 2020; Shneiderman, 2020).

La interpretación humana requiere tiempo y conocimiento. El marco de colaboración de Litvinenko permite reconocer saber situado que una entrada breve puede omitir; la investigación organizativa de Fox y Solihin aporta experiencias de implementación. La síntesis utiliza ambos antecedentes para preguntar qué recursos sostienen revisión, manteniendo sus diseños y límites. Una institución puede limitar temporalmente una función mientras reúne capacidades. Esa decisión necesita motivos y condiciones de reconsideración, evitando que el límite se convierta en una declaración indefinida sin aprendizaje o seguimiento (Fox & Solihin, 2026; Litvinenko, 2026; Tabassi, 2023).

EVALUAR TAMBIÉN AUSENCIA DE MEJORA

El seguimiento debe registrar si una medida funciona, si carece de efecto o si crea obstáculos nuevos. Una mejora en clasificación puede coexistir con una ruta de apelación inaccesible. Una corrección pública puede llegar a algunos destinatarios y dejar otros expuestos. La propuesta exige observar estas diferencias en un periodo y población definidos. Las percepciones aportan información de experiencia y deben relacionarse con acciones y consecuencias. El capítulo ofrece categorías de evaluación y no resultados propios; afirmar eficacia requiere una implementación con comparación y criterios pertinentes (Cabero-Almenara et al., 2026; Chan & Hu, 2023).

La responsabilidad frente al daño se concreta en preservar facultades para comprobar, elegir y corregir. La tecnología puede facilitar información y amplificar perjuicios según uso y contexto. La metáfora de la ira orienta la investigación hacia esa cadena, manteniendo diferencias de intención y resultado. El criterio propuesto combina prevención delimitada, atención pertinente y reparación revisable. Su contribución conceptual consiste en hacer visible qué actor puede modificar cada consecuencia y qué recursos necesita, para que una declaración de cuidado se traduzca en acciones que puedan examinarse (Autio et al., 2024; GFT-CELAM, 2026).

RESPONDER IMPLICA FACULTADES Y TIEMPOS DISTINTOS

La respuesta a un daño requiere coordinar actuaciones que tienen objetos y tiempos diferentes. Conservar evidencia puede ayudar a revisar el caso; limitar circulación puede reducir exposición; ofrecer apoyo atiende consecuencias; corregir decisiones evita que la falsificación siga orientando actuaciones institucionales. Cada medida necesita responsables con capacidad de actuar y debe ajustarse a lo que se sabe, conservando la posibilidad de revisión. La reparación tampoco se reduce a retirar contenido si persisten asociaciones, restricciones o pérdida de oportunidades. Una institución educativa debe explicar cómo se escucha al afectado y qué condiciones permiten recuperar participación sin imponer nuevas cargas innecesarias. Esta propuesta analítica exige evaluación situada antes de presentarse como procedimiento eficaz. Su utilidad consiste en hacer visibles facultades y obligaciones que una respuesta centrada exclusivamente en detectar origen artificial podría omitir, relacionando prevención, atención y aprendizaje institucional con consecuencias observables para las personas (Raji et al., 2020; Suresh & Gutttag, 2021; Tabassi, 2023; UNESCO, 2021).

APLICACIÓN ILUSTRATIVA: UNA INSTRUCCIÓN FALSA CON CONSECUENCIAS ACADÉMICAS

Una institución recibe reportes sobre un audio que aparenta proceder de una autoridad y anuncia un cambio en una fecha de evaluación. El caso es hipotético y no describe un incidente real. Algunas personas manifiestan haber actuado a partir del mensaje. La institución necesita comprobar la instrucción, conocer alcance de circulación y revisar consecuencias, sin concluir de inmediato quién produjo el archivo. Esta definición distingue contenido, atribución y efecto. Permite iniciar una respuesta organizada mientras se conserva la incertidumbre sobre autoría o intención (Tabassi, 2023; Weidinger et al., 2021).

El primer paso consiste en confirmar la instrucción por el canal institucional pertinente. El responsable de la actividad puede establecer qué fecha estaba autorizada y qué comunicaciones existen. Esa comprobación no necesita esperar una atribución técnica completa del audio. Si la instrucción carece de autorización, se comunica esa condición con precisión. La institución identifica dónde debe aclararse y qué acción puede detenerse o revisarse. La propuesta evita que una duda sobre identidad deje vigente una consecuencia que ya puede examinarse mediante documentación del procedimiento (Amershi et al., 2019; Raji et al., 2020).

La conservación de evidencia se limita a materiales necesarios: archivo recibido, fecha, canal y contexto pertinente, conforme a condiciones del procedimiento. Se controla acceso y reproducción. La institución explica a quienes reportan qué información solicita y para qué, sin exigir conversaciones privadas completas. La privacidad se mantiene durante investigación. La propuesta busca evitar que recopilar evidencia amplíe circulación o exposición. Las fichas de datos ayudan a describir composición y finalidad; el diseño ético del aprendizaje aporta criterios para minimizar y ofrecer participación comprensible (Geburu et al., 2021; Lachheb et al., 2023).

El equipo competente examina procedencia en la medida disponible. Puede contrastar registros y analizar información técnica pertinente, conservando límites de cada método. Las credenciales de origen, si existen, se interpretan según su función y versión; su ausencia no

prueba falsedad. La conclusión distingue qué hallazgo se estableció y qué hipótesis permanece. La institución evita atribuir el archivo a una persona sin evidencia suficiente. Este cuidado permite corregir una instrucción comprobada y mantener abierto otro procedimiento de investigación sin confundir sus resultados (Autio et al., 2024; C2PA, 2025).

La comunicación identifica la fecha autorizada, la condición del mensaje y el canal para revisar consecuencias. Debe resultar accesible a quienes pudieron recibirlo. Se propone observar dónde circuló según información disponible y publicar la aclaración en espacios pertinentes, sin afirmar un alcance total desconocido. Una rectificación extensa puede omitir la acción que el destinatario necesita. Las guías de interacción ayudan a revisar claridad y oportunidad de corrección. La comunicación conserva evidencia localizable y evita reproducir innecesariamente la pieza para demostrar su condición (Amershi et al., 2019; Cortina, 2024; Shneiderman, 2020).

La revisión de oportunidades identifica qué personas actuaron bajo la instrucción y qué consecuencia puede examinarse. El procedimiento necesita reglas pertinentes y posibilidad de aportar información. La institución evita asumir que todas recibieron el mismo mensaje o que una explicación individual carece de valor. La propuesta distingue aclaración general y revisión de un caso. El resultado puede corregir un registro o habilitar otra oportunidad conforme a sus obligaciones. Estas decisiones requieren evidencia situada; el capítulo no establece una reparación universal ni una garantía de que toda consecuencia será reversible (Raji et al., 2020; UNESCO, 2021).

La atención considera carga de reclamar. Un formulario puede ser difícil de utilizar por conectividad, horario o lenguaje, y la institución debe ofrecer apoyos que conserven la finalidad. Los antecedentes regionales y de brechas digitales orientan preguntas sin representar datos del caso. Se propone observar si quienes necesitan revisión identifican el canal y reciben una respuesta pertinente. La evaluación mantiene esos indicadores separados de una tasa de falsificación o de un diagnóstico individual. La accesibilidad del procedimiento contribuye a que una facultad formal se vuelva practicable (González-Chouciño & Ruiz-Callado, 2023; Holmes et al., 2022; Kornhaber et al., 2026).

La investigación de responsabilidades sigue una cadena distinta de la revisión de consecuencias. Los documentos de las llamadas de New Hampshire muestran que generación, identificación y transmisión pueden implicar actores diferentes. Su función es un antecedente histórico situado, sin trasladar sanciones o conclusiones penales al caso educativo. La institución identifica qué información puede obtener legítimamente y qué competencia corresponde a cada procedimiento. La atribución necesita evidencia y garantías, mientras cuidado y corrección pueden actuar sobre hallazgos que ya resultan suficientes (FCC, 2024a, 2024b; U.S. Copyright Office, 2024).

La revisión organizativa pregunta qué permitió que la instrucción adquiriera autoridad. Puede examinar si existían canales claros, si una fecha oficial estaba localizable y si el recorrido de confirmación funcionó. Estas preguntas no presuponen una falla concreta antes de estudiarla. La propuesta también incluye a trabajadores y estudiantes para identificar obstáculos y alternativas. Los antecedentes de voz laboral y conocimiento situado permiten considerar recursos y experiencia. Un hallazgo puede modificar comunicación o procedimiento, con un responsable y una fecha de revisión acordados según condiciones institucionales (Doellgast et al., 2026; Fox & Solihin, 2026; Litvinenko, 2026).

El seguimiento debe registrar acciones y efectos conocidos. Interesa si la aclaración llegó a destinatarios pertinentes, si se revisaron oportunidades y qué dificultades permanecen. Retirar una pieza o publicar un comunicado representan actividades, y sus consecuencias necesitan observación. La institución puede encontrar ausencia de mejora y ajustar la respuesta. También debe evitar recopilar información adicional que no contribuye a la finalidad. La auditoría relaciona registro y decisión; su utilidad requiere que el equipo disponga de autoridad y recursos para modificar medidas cuando resultan insuficientes (Lachheb et al., 2023; Raji et al., 2020; Tabassi, 2023).

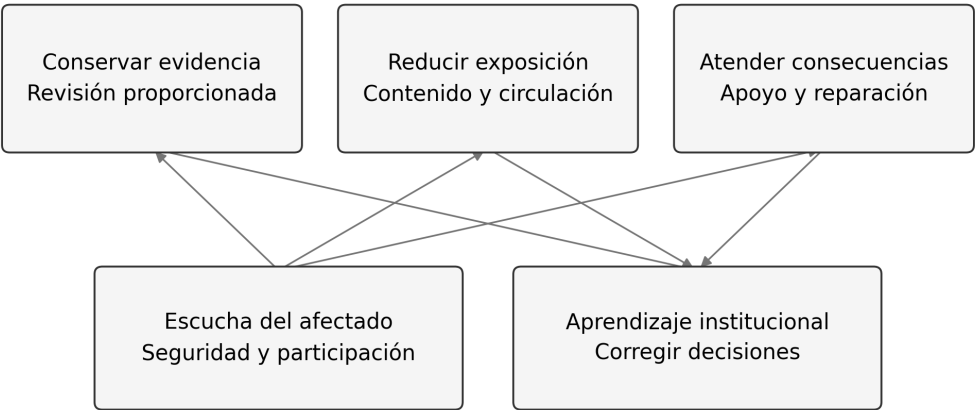
El caso muestra una secuencia donde comprobar, cuidar y reparar conservan objetivos propios y se coordinan. Su eficacia permanece pendiente de implementación evaluada. Una investigación posterior debería definir población, periodo y criterios de comparación, registrar cargas y efectos adversos, y conservar participación. La contribución conceptual consiste en identificar qué evidencia permite cada decisión y qué pregunta sigue abierta. La responsabilidad adquiere contenido cuando una institución puede explicar y revisar esa relación, en lugar de confiar toda la respuesta a una detección técnica o a una declaración de uso ético (Floridi & Cows, 2019).

La evaluación podría contrastar recorridos de respuesta con recursos diferentes,

manteniendo obligaciones comunes de cuidado y explicación. Interesaría observar si disponer de una vía rápida para confirmar instrucciones reduce consecuencias evitables, mientras una investigación más extensa continúa sobre atribución. También convendría registrar casos donde la confirmación no resuelve el problema o añade demoras. Este contraste necesita evitar que el tiempo de cierre sea el único criterio: resolver pronto con una conclusión insuficiente puede crear otro perjuicio. La calidad de evidencia y las oportunidades efectivamente revisadas deben permanecer visibles junto con duración y carga del procedimiento.

La Figura 18 organiza actuaciones que pueden complementarse sin reemplazarse. Conservar evidencia permite revisión; reducir exposición atiende contenido y circulación; ofrecer apoyo responde a consecuencias. La participación del afectado aporta información y exige condiciones de seguridad, mientras el aprendizaje institucional debe cambiar decisiones pertinentes. El esquema conserva la reparación como un problema más amplio que retirar una pieza y permite discutir quién puede actuar, con qué recursos y qué resultados deberían observarse para valorar la respuesta (Raji et al., 2020; Tabassi, 2023; UNESCO, 2021).

Figura 18. Objetos de una respuesta institucional al daño algorítmico



Nota. Elaboración conceptual sobre actuaciones con objetos diferentes. La escucha del afectado y el aprendizaje institucional atraviesan la respuesta. Es una propuesta analítica que requiere evaluación; no acredita eficacia de implementación propia. Fuente: (Raji et al., 2020; Tabassi, 2023; UNESCO, 2021).

Tabla 8. Objetos y obligaciones de una respuesta al daño

Actuación	Objeto de la respuesta	Condición pertinente
Conservar evidencia	Permitir revisión del caso	Privacidad y proporcionalidad
Reducir exposición	Contenido y circulación	Facultad y rapidez de intervención
Ofrecer apoyo	Consecuencias para el afectado	Escucha y seguridad
Corregir decisiones	Efectos institucionales persistentes	Responsables y recursos

CONCLUSIONES

La ira tecnológica se estudia mediante mecanismos de daño y cadenas de responsabilidad. Distinguir error, intención y discriminación permite intervenir donde corresponde. La escala automática puede agravar un conflicto, pero el control debe observar también objetivos, relaciones de poder y capacidad institucional de corregir. El cuidado se vuelve efectivo cuando ofrece procedimientos que previenen exposición, atienden a las personas y permiten reparar decisiones.

La responsabilidad se debilita si quienes utilizan los sistemas abandonan sus capacidades de juicio. La siguiente discusión examina la pereza como renuncia a aprender, pensar o decidir cuando la asistencia ofrece una salida inmediata. Conservar criterio resulta necesario para aprovechar beneficios y para detectar las situaciones en que una herramienta contribuye al daño que este capítulo ha estudiado.

La revisión identifica una cadena en la que capacidades de IA pueden amplificar perjuicios o apoyar actividades pertinentes. El efecto depende de objetivo, datos, salida, distribución y

decisión. Intención, mecanismo y consecuencia conservan exigencias diferentes de evidencia. Una institución puede corregir una instrucción o atender exposición mientras sigue investigando atribución. Esta separación permite actuar con razones y límites, evitando convertir toda falla en agresión deliberada o eliminar responsabilidad porque el daño carece de una intención manifiesta (Selbst et al., 2019; Suresh & Guttag, 2021; Weidinger et al., 2021).

Los estudios de sesgo muestran mecanismos distintos de desigualdad, y las evaluaciones de mensajes y toxicidad conservan tareas y medidas. La síntesis mantiene su alcance para que una crítica no transforme posibilidad técnica en prevalencia general. La investigación de confianza y automatización añade condiciones de supervisión material: tiempo, información, práctica y autoridad. La presencia de una persona resulta relevante y necesita esas capacidades para modificar una consecuencia. También se conserva evidencia sobre rechazo excesivo de ayuda, evitando formular cuidado como desconfianza uniforme (Bainbridge, 1983; Buolamwini & Gebru, 2018; Dietvorst et al., 2015; Gehman et al., 2020; Obermeyer et al., 2019).

La desinformación exige comprobar contenido y estudiar circulación. Una salida convincente puede ser verdadera o falsa; el origen declarado no determina exactitud. La procedencia técnica aporta un insumo y no certifica consentimiento o verdad. La confianza pública requiere conservar posibilidades de confirmar registros legítimos y retirar atribuciones falsas. Esa doble obligación evita que la existencia de falsificaciones se convierta en una razón para descartar toda evidencia. Los procedimientos deben comunicar qué se pudo comprobar y qué permanece incierto (Chesney & Citron, 2019; C2PA, 2025; Goldstein et al., 2023; Spitale et al., 2023).

La respuesta institucional integra prevención, atención y reparación, con objetivos y responsables. Prevenir delimita funciones; atender reduce exposición y ofrece apoyo; reparar revisa oportunidades y registros. Estas acciones requieren recursos y deben observarse en consecuencias, además de actividad. La privacidad y accesibilidad permanecen condiciones durante el proceso. La aplicación propuesta ilustra coordinación y límites, sin resultados propios. Afirmar eficacia exige una implementación evaluada que incluya ausencia de mejora y efectos adversos, con participación de quienes utilizan o cuestionan el procedimiento (Holmes et al., 2022; Lachheb et al., 2023; Raji et al., 2020; Tabassi, 2023).

La agenda posterior debe estudiar distribución de errores, autoridad de señales automatizadas y capacidad real de reclamar. También conviene comparar qué correcciones alcanzan destinatarios y qué cargas crean los canales de respuesta. Las fuentes organizativas aportan preguntas situadas, sin garantizar transferencia a toda institución. El capítulo no estima prevalencia ni establece consecuencias jurídicas de casos particulares. Su aporte conceptual relaciona evidencia y facultad de modificar, para que responsabilidad y cuidado puedan examinarse mediante decisiones y resultados (Cui, 2026; Doellgast et al., 2026; Fox & Solihin, 2026).

La metáfora de la ira resulta útil cuando orienta hacia aquello que puede cambiarse: objetivos, uso, circulación y procedimientos que vuelven una salida decisiva. Las seis facultades propuestas en el libro permiten preguntar si las personas pueden comprobar, elegir y corregir dentro de esa cadena. La corrección atraviesa todos los capítulos y requiere poder institucional. Conservarla permite aprovechar asistencia y limitar usos que producen consecuencias injustificadas, con una deliberación abierta a nueva evidencia y una responsabilidad humana explícita (Autio et al., 2024; Cortina, 2024; GFT-CELAM, 2026; UNESCO, 2021).

Declaración de uso de inteligencia artificial

Se utilizaron herramientas de inteligencia artificial generativa como apoyo en los procesos de organización, redacción y revisión del manuscrito. Su utilización se limitó a funciones de asistencia editorial y analítica, bajo supervisión y validación permanente de los autores. Las propuestas, interpretaciones y escenarios desarrollados en el texto corresponden a elaboraciones de carácter analítico y no constituyen resultados derivados de procesos de implementación empírica. Las herramientas de inteligencia artificial empleadas no reciben atribución de autoría ni sustituyen las fuentes científicas, documentales y académicas citadas. La selección de las fuentes, la interpretación de la evidencia, la construcción argumentativa, la revisión final del contenido y la decisión de someter el manuscrito a publicación son responsabilidad exclusiva de los autores.

REFERENCIAS

- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. En *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300233>
- Ares-Ferreirós, M., de la Cruz-Campos, J. C., Morodo Horrillo, A., & Robert Guillén, S. (2026). Buenas prácticas en el uso de la IA generativa en las universidades. Un análisis documental. *Innoeduca International Journal of Technology and Educational Innovation*, 12(1), 5–20. <https://doi.org/10.24310/ijtei.121.2026.22152>
- Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.600-1>
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
- Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2), 220–239. <https://doi.org/10.1002/bdm.2155>
- Cabero-Almenara, J., Barroso-Osuna, J., Loaiza Aguirre, M. I., & Palacios-Rodríguez, A. (2026). Artificial intelligence in higher education: Cognitive, emotional, behavioral and ethical attitudes of Latin American students. *Journal of New Approaches in Educational Research*, 15, Article 13. <https://doi.org/10.1007/s44322-026-00063-2>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), Article eaec8352. <https://doi.org/10.1126/science.aec8352>
- Cheng, Q., Dai, Y., Liu, X., & Peng, S. (2026). The trust crisis in artificial intelligence: AI hallucinations and human-AI collaboration. *Technology in Society*, 86, Article 103286. <https://doi.org/10.1016/j.techsoc.2026.103286>
- Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820. <https://doi.org/10.15779/Z38RV0D15J>
- Coalition for Content Provenance and Authenticity. (2025). *C2PA technical specification* (Versión 2.2). https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html
- Cortina, A. (2024). *¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada*. Paidós.
- Crocchio, E., Ballesio, E., Broccardo, L., & Alghafes, R. (2026). Welcome to the dark side: Unveiling the potential risks related to the implementation of Artificial Intelligence. *Journal of Risk Research*. Publicación anticipada en línea. <https://doi.org/10.1080/13669877.2025.2611954>
- Cui, Y. (2026). Artificial intelligence and human assessment consistency and bias in higher education academic writing evaluation. *Discover Artificial Intelligence*, 6, 559. <https://doi.org/10.1007/s44163-026-01286-0>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Doellgast, V., Kämpf, T., & Langes, B. (2026). Building worker voice and power in AI decisions: Three cases in the German ICT industry. *Work and Occupations*. Publicación anticipada en línea. <https://doi.org/10.1177/07308884251412886>
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 625–630. <https://doi.org/10.1038/s41586-024-07421-0>
- Federal Communications Commission. (2024a). *In the matter of Lingo Telecom, LLC: Order and consent decree* (DA 24-790). <https://docs.fcc.gov/public/attachments/DA-24-790A1.pdf>
- Federal Communications Commission. (2024b). *FCC issues \$6 million fine for New Hampshire deepfake robocalls*. <https://docs.fcc.gov/public/attachments/DOC-405811A1.pdf>
- Filippelli, S., Popescu, I. A., Verteramo, S., Tani, M., & Corvello, V. (2026). Generative AI and employee well-being: Exploring the emotional, social, and cognitive impacts of adoption. *Journal of Innovation & Knowledge*, 11, 100844. <https://doi.org/10.1016/j.jik.2025.100844>
- Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Fox, G., & Solihin, Y. (2026). Exploring organizational implementation of generative AI: Navigating benefits and safety risks. *Journal of Decision Systems*, 35(1), 2649940. <https://doi.org/10.1080/12460125.2026.2649940>
- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198237907.001.0001>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021).

- Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. En *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 3356–3369). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.findings-emnlp.301>
- Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative language models and automated influence operations: Emerging threats and potential mitigations* [Prepublicación]. arXiv. <https://arxiv.org/abs/2301.04246>
- González-Chouciño, M. A., & Ruiz-Callado, R. (2023). Analysis of the third digital divide in relation to digital socialization itineraries among university students in Uruguay. *Societies*, 13(12), Article 252. <https://doi.org/10.3390/soci13120252>
- Grupo de Trabajo de Frontera Tecnológica del Celam. (2026). *Inteligencia artificial y desarrollo humano integral*. Consejo Episcopal Latinoamericano y Caribeño. <https://repositorio.celam.org/s/repositorio/item/3387>
- Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 32, 504–526. <https://doi.org/10.1007/s40593-021-00239-1>
- International Committee of the Red Cross. (2021). *ICRC position on autonomous weapon systems*. <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kornhaber, R., Bragg, S., Matheson, A., & Irwin, P. (2026). Qualitative inquiry into generative artificial intelligence impact on rural and regional higher education. *Australian and International Journal of Rural Education*, 36(2), 1–23. <https://doi.org/10.47381/aijre.v36i2.812>
- Lachheb, A., Abramenska-Lachheb, V., Moore, S., & Gray, C. (2023). The role of design ethics in maintaining students' privacy: A call to action to learning designers in higher education. *British Journal of Educational Technology*, 54, 1653–1670. <https://doi.org/10.1111/bjet.13382>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), Article 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. En *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.229>
- Litvinenko, A. (2026). The knowledge collaboration framework: A review of artificial intelligence in organisational learning and knowledge management. *Journal of Decision Systems*, 35(1), 2616703. <https://doi.org/10.1080/12460125.2026.2616703>
- Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. En *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376727>
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.173>
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. En *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, Article 100041. <https://doi.org/10.1016/j.caeai.2021.100041>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. En *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372873>
- Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. En *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287598>
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International*

- Journal of Human-Computer Interaction*, 36(6), 495–504. <https://doi.org/10.1080/10447318.2020.1741118>
- Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances*, 9(26), Article eadh1850. <https://doi.org/10.1126/sciadv.adh1850>
- Suresh, H., & Gutttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. En *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, Article 17. Association for Computing Machinery. <https://doi.org/10.1145/3465416.3483305>
- Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- U.S. Copyright Office. (2024). *Copyright and artificial intelligence. Part 1: Digital replicas*. <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf>
- UNESCO. (2021). *Recomendación sobre la ética de la inteligencia artificial*. <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, Article 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). *Ethical and social risks of harm from language models* [Prepublicación]. arXiv. <https://doi.org/10.48550/arXiv.2112.04359>
- World Health Organization. (2021). *Ethics and governance of artificial intelligence for health*. <https://www.who.int/publications/i/item/9789240029200>
- Xie, M., Xu, J., Liu, Y., Su, H., Long, Q., & Zou, Y. (2026). Work autonomy mediates associations between medical AI readiness and well being in a three wave nurse study. *Scientific Reports*, 16, 23474. <https://doi.org/10.1038/s41598-026-58212-8>
- Yue, X., Zhang, Y., Kang, M., Hu, H., & Li, C. (2026). The impact of AI-integrated purchasing management on managers' well-being: A dual-pathway view. *Transportation Research Part E Logistics and Transportation Review*, 212, 104942. <https://doi.org/10.1016/j.tre.2026.104942>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16, Article 39. <https://doi.org/10.1186/s41239-019-0171-0>