Revista Científica del Instituto Nacional de Medicina Legal y Ciencias Forenses de Colombia

https://revistasforenses.medicinalegal.gov.co/index.php/an/about

Artículo original

Inteligencia artificial generativa en contextos forenses: lineamientos éticos, salvaguardas de derechos humanos y estándares de uso responsable

Generative Artificial Intelligence in Forensic Contexts: Ethical Guidelines, Human Rights Safeguards, and Standards for Responsible Use

Aurelio Israel Coronado Mares[1], Karla Valeria Baltazar Torres[2], Víctor Alejandro Nodal Silva[3]

[1] Doctor en Psicología, Equipo Independiente de Investigación en Ciencias Forenses y Derechos Humanos, México.
[2] Maestra en Ciencias Forenses, Equipo Independiente de Investigación en Ciencias Forenses y Derechos Humanos, México.
[3] Doctor en Derechos Humanos, Equipo Independiente de Investigación en Ciencias Forenses y Derechos Humanos, México.

BY NC SA

Historia del artículo

Recibido: 09/04/2026

Aceptado: 27/08/2026

Antistio, vol. 3, 2026, e5897

Resumen

Este artículo propone lineamientos éticos, salvaguardas de derechos humanos y un estándar mínimo de uso responsable de inteligencia artificial generativa en contextos forenses. Sostiene que el riesgo central se encuentra en el cierre por relato, es decir, textos plausibles que rellenan vacíos, desplazan el rastro verificable y elevan el costo de la contradicción, afectando el debido proceso, el derecho a la prueba y el derecho a la verdad. Con base en un criterio de riesgo-proporcionalidad, se distinguen usos de apoyo y usos con pretensión decisoria, estableciendo requisitos de trazabilidad, respaldo localizable y límites explícitos, además de prohibiciones frente a opacidad, perfilación y datos sensibles. Se plantea un control humano experto y auditable como elemento obligatorio de todo uso de inteligencia artificial generativa para fines forenses.

Palabras claves: ciencias forenses, aprendizaje automático, administración de justicia, ética, trazabilidad.

correspondencia: victor.nodal4177@academicos.udg.mx

Abstract

This article proposes ethical guidelines, human rights safeguards, and a minimum standard for the responsible use of generative artificial intelligence in forensic contexts. It argues that the central risk lies in narrative closure, that is, plausible texts that fill in gaps, displace the verifiable trail, and raise the cost of adversarial challenge, thereby affecting due process, the right to evidence, and the right to the truth. Based on a risk-proportionality criterion, the article distinguishes between supportive uses and uses intended to carry decision-making weight, establishing requirements for traceability, locatable supporting material, and explicit limits, as well as prohibitions concerning opacity, profiling, and sensitive data. It proposes auditable expert human oversight as a mandatory element of any use of generative artificial intelligence for forensic purposes.

Keywords: forensic science, machine learning, administration of justice, ethics, traceability.

Résumé

L'intelligence artificielle générative dans les contextes médico-légaux: lignes directrices éthiques, garanties en matière de droits humains et normes d'utilisation responsable

Cet article propose des lignes directrices éthiques, des garanties en matière de droits humains et une norme minimale d'utilisation responsable de l'intelligence artificielle générative dans les contextes médico-légaux. Il soutient que le risque principal réside dans la fermeture narrative, c'est-à-dire la production de textes plausibles qui comblent les lacunes, font disparaître les traces vérifiables et augmentent le coût de la contradiction, ce qui porte atteinte au respect des garanties procédurales, au droit à la preuve et au droit à la vérité. Sur la base d'un critère de proportionnalité des risques, une distinction est opérée entre les utilisations à titre d'appui et celles ayant une vocation décisionnelle, en établissant des exigences de traçabilité, de fondement localisable et de limites explicites, ainsi que des interdictions concernant l'opacité, le profilage et les données sensibles. Un contrôle humain expert et auditable est proposé comme élément obligatoire de toute utilisation de l'intelligence artificielle générative à des fins médico-légales.

Mots-clés: sciences médico-légales, apprentissage automatique, administration de la justice, éthique, traçabilité.

Resumo

Inteligência artificial generativa em contextos forenses: diretrizes éticas, salvaguardas dos direitos humanos e padrões para o uso responsável

Este artigo propõe diretrizes éticas, salvaguardas dos direitos humanos e um padrão mínimo para o uso responsável da inteligência artificial generativa em contextos forenses. Sustenta que o risco central reside no fechamento narrativo, isto é, textos plausíveis que preenchem lacunas, desviam os rastros verificáveis e aumentam o custo da contradição, afetando o devido processo legal, o direito à prova e o direito à verdade. Com base em um critério de risco-proporcionalidade, distinguem-se usos de apoio e usos com pretensão decisória, estabelecendo requisitos de rastreabilidade, suporte localizável e limites explícitos, além de proibições diante de opacidade, perfilamento e dados sensíveis. Propõe-se um controle humano especializado e auditável como elemento obrigatório de todo uso de inteligência artificial generativa para fins forenses.

Palavras-chave: ciências forenses, aprendizado de máquina, administração da justiça, ética, rastreabilidade.

Introducción

La crisis forense tiende a producir un problema estructural de derechos humanos, en el que el Estado pierde capacidad real para investigar con la debida diligencia y, en ese vacío, el contenido de las carpetas de investigación puede desplazarse desde el rastro verificable hacia una construcción narrativa convincente, pero sin sustento, es decir, hacia lo que "suena" completo, coherente y técnicamente

redactado, conocido en el argot penal como "cuadrar la carpeta de investigación". En ese escenario, la inteligencia artificial (IA) ingresa como una potencia capaz de acelerar dos trayectorias opuestas que, por un lado, pueden fortalecer el trabajo duro de investigación (ordenar información dispersa, detectar inconsistencias, mapear lagunas probatorias, armonizar en debida diligencia y contrastar hipótesis competidoras); pero, por otro, pueden reforzar el cierre por relato, produciendo un texto persuasivo que llena vacíos con plausibilidad, sin ofrecer evidencia suficiente para auditar lo dicho o, en el peor de los casos, producir narrativas soportadas en prejuicios, estereotipos y discursos discriminatorios. De ahí se desprende que, en contextos forenses, el uso de IA generativa exige un anclaje explícito en derechos humanos para impedir que la forma narrativa sustituya la verificación, para mantener la probabilidad práctica de la contradicción (posibilitando trazabilidad, control de fuentes y réplica efectiva) y para evitar un daño diferencial en poblaciones en situación de vulnerabilidad, así como en casos con mayor exposición a sesgos y desigualdades.

Para fines de este artículo, "inteligencia artificial generativa" se entiende como el conjunto de sistemas capaces de producir texto (y, en su caso, síntesis o reformulaciones) con apariencia de coherencia, a partir de instrucciones y de insumos documentales, como parte de un proceso sociotécnico de apoyo a tareas de análisis y redacción (1). Su rasgo críticamente problemático en contextos forenses no es la automatización en sí, sino la capacidad de completar vacíos epistémicos con narrativa persuasiva, generando salidas plausibles que pueden incorporar contenido no verificado o inconsistente, con potencial de producir efectos de credibilidad indebida (2,3). Por ello, el análisis se desarrolla desde escenarios de crisis forense como caso límite, pero los lineamientos y las salvaguardas aquí propuestos son extrapolables a otros contextos forenses, ajustando el nivel de control y trazabilidad según el riesgo de daño y el peso decisorio del producto generado, en clave de proporcionalidad y evaluación de impacto basada en derechos humanos (1).

Asimismo, el presente artículo adopta una perspectiva regional, sustentada en el sistema interamericano de derechos humanos (SIDH). En consecuencia, las categorías de debida diligencia, garantías judiciales, protección judicial, derecho a la verdad, enfoque diferencial y deber estatal de investigar son comprendidas a partir de las obligaciones establecidas en la Convención Americana sobre Derechos Humanos (CADH) y de su desarrollo progresivo en la jurisprudencia de la Corte Interamericana de Derechos Humanos (Corte IDH) (4,5). Las decisiones judiciales y las disposiciones nacionales pueden concretar o ampliar estas obligaciones dentro de cada ordenamiento, pero su examen sistemático excede el objeto de este trabajo, cuyo propósito es proponer un estándar técnico-científico susceptible de aplicación regional y de posterior armonización con el derecho interno de cada Estado. Ante la ausencia de un instrumento interamericano vinculante específicamente dedicado al uso de inteligencia artificial, la propuesta desarrolla una interpretación aplicada de obligaciones convencionales preexistentes frente a los riesgos particulares de la IA generativa, sin sustituir las reglas nacionales sobre prueba, protección de datos, responsabilidad profesional o procedimiento.

El riesgo del uso de inteligencia artificial generativa en contextos forenses

En escenarios de crisis, la investigación criminal tiende a fragmentarse, como resultado, faltan registros que serían esperables, se diluye la cadena de custodia y se multiplican los oficios sin trazabilidad

ni seguimiento. En ese deterioro, la forma puede terminar sustituyendo a la debida diligencia. El resultado es teatro forense, es decir, la apariencia de método sin articulación suficiente del rastro forense y la evidencia (6).

La legibilidad institucional, es decir, la capacidad del sistema para construir e interpretar narrativas, se vuelve especialmente decisiva en crisis. Los sistemas institucionales suelen operar mediante simplificaciones necesarias: clasifican, estandarizan y agregan información para volverla manejable, a fin de construir una teoría del caso escenificable en juicio. Sin embargo, cuando el entorno está degradado, esa legibilidad puede sustituir el lugar del rastro y la evidencia, entonces el documento se convierte en un "hecho operativo", aun si está incompleto. En ese escenario, la IA puede acelerar y soportar la producción de textos plausibles, pero insuficientes en términos probatorios y científicos, los cuales, en la práctica, funcionan como hechos de papel. El problema principal no es el "error" en un sentido estrictamente técnico, sino la suficiencia aparente, en donde aparece un texto verosímil que, por su coherencia interna y su tono, se lee como conclusión.

Desde la perspectiva del acceso a la justicia y el debido proceso, la consecuencia es crítica: sin investigación competente no es posible establecer con claridad si hubo un crimen ni delimitar quién es la víctima (7).

A partir de lo anterior, identificamos elementos recurrentes que conviene describir de forma explícita, en tanto funcionan como modos típicos de errores cuando la IA generativa se incorpora sin controles suficientes y sin una lectura metodológica del expediente:

1. Sustitución de diligencias por síntesis: ocurre cuando la herramienta se usa para "cerrar" tareas que en realidad exigen investigación (verificación de fuentes, contraste de versiones, identificación de vacíos, reconstrucción de secuencias, etc.). La síntesis puede ser útil como apoyo, pero se vuelve dañina si reemplaza diligencias mínimas.

2. Autoridad por estilo y contaminación entre planos: se presenta cuando la credibilidad del contenido no deriva del rastro verificable, sino del tono técnico, la fluidez argumentativa o el hecho mismo de haber sido producido "con IA". El estilo puede crear una ilusión de rigor, y el dispositivo funcionar como sello de exactitud y, sin advertirlo, contaminar niveles interpretativos que deben mantenerse diferenciados, por ejemplo, hechos e hipótesis, descripción e interpretación, dato documentado e inferencia, evidencia y explicación.

3. Aumento de dificultad en la contradicción: consiste en elevar la dificultad práctica de rebatir lo afirmado. Un texto altamente coherente puede obligar a invertir más recursos para separar datos de inferencias, rastrear orígenes y desmontar afirmaciones, lo que genera asimetría.

4. Daño diferencial por estereotipos o variables sustitutas: se refiere a impactos desiguales cuando el sistema reproduce patrones estereotípicos o utiliza correlaciones con atributos sensibles (p.

ej., territorio, escolaridad, ocupación, lenguaje o historial penal). El error no siempre es explícito; suele expresarse como atribuciones implícitas, sospechas normalizadas o lecturas sesgadas de credibilidad, intencionalidad o peligrosidad.

5. Retrospección y evidencia digital abierta: las IA generativas construyen sus salidas a partir de patrones aprendidos en bases de datos y corpus previamente curados, de modo que la coherencia del resultado no proviene solo de los insumos del caso, sino también de decisiones de selección, exclusión y jerarquización realizadas en etapas previas de entrenamiento o ajuste. Esa curaduría puede incorporar sesgos, omisiones sistemáticas o criterios de representación que entrañan riesgos relevantes para la protección de derechos humanos, de modo que, sin formación especializada, un actor jurídico puede no ser capaz de distinguir cuándo el texto refleja evidencia verificable del expediente o cuándo está siendo moldeado por el modelo en cuestión y las fuentes a las que tiene acceso. En esos contextos, la IA puede reforzar desigualdades preexistentes: quien controla recursos controla narrativas (8).

6. Integración sin guías: debido a la masificación de la herramienta, es fácil inferir que la adopción de IA en el campo de la justicia ya ocurre a gran escala, sin embargo, la disponibilidad de guías institucionales y de capacitación suele ir muy por detrás de la práctica cotidiana de uso. Una respuesta compatible con los derechos humanos requiere gobernanza proporcional, esto es, controles ajustados al riesgo y al peso decisorio del producto, así como formación especializada, de manera que lo generado pueda ser examinado, discutido y contradicho con trazabilidad suficiente.

Para fines pedagógicos, el siguiente ejemplo hipotético permitirá ilustrar estos fenómenos: una persona operadora del sistema de justicia investiga una desaparición. En la carpeta existen tres entrevistas con versiones parcialmente contradictorias, el análisis de telefonía aún no concluye y falta practicar una diligencia para verificar la última ubicación conocida de la víctima. Con el propósito de presentar un avance de la investigación, la persona operadora solicita a una herramienta de inteligencia artificial generativa que elabore un informe con la información disponible. La herramienta produce un documento claro, bien estructurado y aparentemente completo. En él, las contradicciones entre las entrevistas desaparecen porque son integradas en una sola narrativa coherente; la ausencia del análisis telefónico deja de percibirse como un vacío relevante, y la diligencia pendiente ya no aparece como una necesidad para continuar investigando, sino como un detalle secundario. Quien lee el documento puede concluir que la investigación se encuentra prácticamente consolidada, aunque el expediente todavía presenta incertidumbres sustanciales. El problema no consiste en que la herramienta haya inventado hechos, sino en que la coherencia de la redacción produce una apariencia de suficiencia que desplaza la atención desde aquello que aún debe investigarse hacia un relato que parece explicar satisfactoriamente lo ocurrido. En ese momento, la síntesis deja de ser un apoyo metodológico y comienza a sustituir la investigación misma. Es precisamente este desplazamiento del rastro verificable hacia una narrativa de suficiencia aparente lo que en este trabajo se denomina cierre por relato.

Perspectiva de derechos humanos: qué exige y cómo previene riesgos

En un sistema adversarial, la perspectiva de derechos humanos funciona como un estándar transversal de operación para cualquier actor involucrado: exige contradicción real, control del poder punitivo, debida diligencia y prevención, y reparación del daño con enfoque diferencial. Su finalidad es evitar que la forma sustituya al rastro y que la desigualdad de recursos termine convertida en desigualdad de facto. En términos más amplios, este estándar se sostiene en un andamiaje transdisciplinar que busca coherencia entre supuestos de conocimiento, comprensión de la realidad y criterios prácticos de intervención, de modo que la metodología conserve sentido para la tutela, la promoción y el reconocimiento de la dignidad humana (9).

Desde esa lógica, se parte de la premisa operativa de que las personas son titulares de derechos con capacidad jurídica y social para exigir el cumplimiento de deberes estatales. Este punto es clave para el uso de IA en contextos forenses, en donde no basta con que el producto sea "útil" o "eficiente"; debe ser compatible con obligaciones de debida diligencia, control de poder y garantías procesales.

En términos metodológicos, esta perspectiva se traduce en los principios rectores aplicables a cualquier operador, tales como presunción de inocencia, derecho a la prueba, defensa adecuada y contradicción real, debido proceso, duda razonable, elevación de estándares cuando el Estado pretende restringir derechos, carga de la prueba en el Estado, y una orientación al control del daño y del abuso de poder.

Traducido a exigencias operativas en el uso de herramientas de IA, esta perspectiva impone, como mínimo, varias condiciones, enunciando (mas no limitando) las que a continuación se presentan:

1. Debida diligencia: ninguna síntesis puede sustituir diligencias; la IA, en todo caso, solo puede apoyar, ordenar o abrir rutas, pero no "cerrar" por sí misma.

2. Garantía de derechos procesales y contradicción real: todo enunciado que pretenda tener peso debe poder revisarse, es decir, debe ser claro qué se afirma, dónde está el respaldo, qué límites tiene y qué información falta.

3. Enfoque diferencial e interseccional: cuando existe riesgo de impacto desigual, deben activarse salvaguardas reforzadas para evitar daño diferencial y reproducción de sesgos, particularmente en redacciones y análisis realizados con IA y que se utilicen para interpretación.

4. Derecho a la verdad: en la crisis forense, el cierre por relato produce una forma de amnesia institucional, porque se normaliza que "no hay más", aunque falten registros esperables y existan alternativas plausibles no exploradas. La verdad pública depende del rastro verificable y también del reconocimiento explícito de aquello que no se sabe. Bajo estándares de trazabilidad y tutela humana, la IA debe convertirse en un instrumento que documente la verdad.

En esa clave, documentar vacíos constituye un acto de honestidad institucional y un cumplimiento de obligación que permite identificar rutas de diligencia, evita cierres falsos y preserva la posibilidad de reparación. De ahí que el estándar deba favorecer tres conductas básicas: 1) preservar y describir el rastro disponible, 2) declarar límites de inferencia, y 3) dejar constancia de vacíos relevantes que impiden concluir.

La misma lógica se proyecta en los mínimos técnicos vinculados con la integridad de la evidencia y con negligencia. En este contexto, la negligencia puede consistir en incumplir estándares o protocolos por acción u omisión, afectando directamente la posibilidad de contradicción (7). En términos probatorios, la confiabilidad y el alcance de los métodos periciales deben justificarse y comunicarse con límites y, cuando aplique, con posibles tasas de error (10,11). Por ello, cuando un enunciado pretende tener peso decisivo, debe formularse de manera revisable y responder, con trazabilidad suficiente, a preguntas básicas: qué se afirma, dónde está el respaldo, qué límites tiene, qué falta y qué salvaguardas se aplicaron (12).

Desde la perspectiva del derecho interamericano de los derechos humanos, la utilización de herramientas de IA no modifica el régimen de atribución de responsabilidades del Estado. Cuando una decisión estatal apoyada en IA contribuye a una investigación deficiente, a una imputación carente de sustento o a un dictamen construido sobre inferencias no verificables, la cuestión jurídica central no consiste en determinar si el error fue "humano" o "tecnológico", sino en establecer si el Estado incumplió su deber de organizar todo su aparato institucional para garantizar una investigación realizada con debida diligencia y compatible con las garantías judiciales. Conforme a la jurisprudencia interamericana, los Estados responden por los actos y las omisiones de sus agentes, así como por la ausencia de controles suficientes para prevenir afectaciones a los derechos protegidos por la CADH. En consecuencia, la incorporación de IA generativa no atenúa los deberes estatales de supervisión, motivación, trazabilidad y control humano experto; por el contrario, incrementa la exigencia de demostrar que toda conclusión con efectos jurídicos permaneció sometida a revisión crítica, fue susceptible de contradicción y conservó un respaldo probatorio verificable.

Control humano experto como garantía de impugnabilidad y auditabilidad externa

En la justicia penal, el criterio rector es el binomio riesgo-proporcionalidad, es decir, a mayor riesgo de afectación de los derechos humanos, mayor exigencia de controles; y, cuando ese riesgo no es mitigable, corresponde prohibir el uso de IA en funciones con peso decisorio (13,14). Para hacerlo operable sin añadir burocracia innecesaria, pueden emplearse marcos técnicos de gestión de riesgos que priorizan trazabilidad, evaluación y gobernanza a lo largo del ciclo de vida de los sistemas de IA (15). En el caso de sistemas generativos, los riesgos asociados a la producción de texto plausible, a errores no detectados y a la degradación de la evidencia exigen controles reforzados y documentación explícita de límites (1). Como marco global, la recomendación de la Unesco sobre ética de la IA establece principios de supervisión humana, transparencia, equidad y rendición de cuentas compatibles con este estándar (16).

Como respuesta a estas necesidades, la propuesta es una revisión y control basado en un análisis metodológico por niveles. La idea central es que no toda tarea requiere el mismo nivel de control. La intensidad del control se activa por función (apoyo o peso decisivo) y por riesgo (daño diferencial, datos sensibles, opacidad). Este esquema no presupone que todas las instituciones cuenten con la misma capacidad operativa. La intensidad de los controles debe ajustarse al principio de riesgo-proporcionalidad y a las capacidades institucionales disponibles, sin renunciar a los mínimos compatibles con los derechos humanos. Así, en tareas de apoyo o de bajo riesgo puede resultar suficiente la revisión por la persona responsable del producto; cuando el resultado pretenda adquirir peso decisorio o exista un riesgo elevado de afectación de derechos fundamentales, deberán incorporarse controles metodológicos adicionales acordes con las posibilidades de la institución. El estándar, por tanto, exige garantizar que el nivel de supervisión sea proporcional al riesgo jurídico y que ninguna conclusión con efectos relevantes quede exenta de control humano experto.

En este sentido, lo propuesto en el protocolo institucional de caracterización victimológica en graves afectaciones contra los derechos humanos (17) puede servir como apoyo metodológico porque ordena el trabajo en tres dimensiones y una secuencia de pasos orientados a fortalecer la vocación probatoria sin sustituir diligencias. Este modelo, aplicado al uso de IA, se centra en el método y la trazabilidad del proceso. A continuación, se presenta dicho modelo:

a. Dimensión técnica (insumos, resguardo, trazabilidad)

El objeto de esta dimensión es controlar qué entra, qué se excluye y qué se protege para evitar el uso de insumos defectuosos (datos sin origen, sin contexto de obtención o con información sensible indebidamente expuesta, etc.). La dimensión técnica fija estándares mínimos, en los que se incluye la delimitación de datos, recomendando que antes de usar IA generativa, se defina qué registros son válidos para la tarea (actas, entrevistas, peritajes, etc.) y se prohíba alimentar la herramienta con mezclas no distinguibles (hecho, versión, opinión, rumor) cuando el producto pueda circular a carpeta o dictamen. También se exige registrar el origen y las condiciones de obtención, es decir, para cada insumo relevante se anota cómo se obtuvo y qué limitaciones presenta (p. ej., copia parcial, fuente secundaria, transcripción incompleta, edición). En evidencia digital abierta, el umbral es más estricto, sin rastro del proceso, el producto no puede pesar (18). A ello se suma el resguardo de información sensible, para lo que deben activarse salvaguardas reforzadas cuando hay niñas, niños y adolescentes, discapacidad, comunidades indígenas, víctimas, datos biométricos, integrantes de la comunidad LGBT+, personas defensoras de derechos humanos, información íntima, etc., siendo el criterio rector minimizar el daño usando solo lo necesario para el propósito legítimo (19). Finalmente, se requiere trazabilidad práctica: registrar qué versión del texto se generó, con qué insumos y con qué límites.

b. Dimensión metodológica (preguntas, hipótesis, contraste, límites)

Esta dimensión se centra en la lógica con la que se investiga el caso. La IA solo resulta aceptable como apoyo si fortalece el método (formular preguntas, organizar evidencia, proponer rutas de diligencia

y detectar contradicciones). Por ello, el primer mínimo es el planteamiento del problema, es decir, convertir solicitudes vagas o sesgadas en preguntas resolubles; por ejemplo, en lugar de pedir que "la herramienta determine", resulta más útil preguntar qué evidencia existe, qué falta y qué puede concluirse con límites. Un segundo mínimo es trabajar con hipótesis alternativas. Cuando el texto pretenda peso decisivo, se exige al menos dos hipótesis competidoras; la herramienta puede ayudar a enumerarlas, pero el operador debe justificar por qué una se sostiene mejor y qué evidencia faltante movería la conclusión. En tercer mínimo es la contrastación y consistencia. Aquí, el método obliga a buscar convergencias entre fuentes y detectar discordancias. A esto se agregan límites explícitos, como que toda interpretación debe declarar qué no puede concluirse, es decir, los límites naturales de la inferencia. Por último, los vacíos relevantes deben describirse indicando qué falta, por qué importa y qué diligencia lo resolvería; documentar vacíos es una obligación emanada del principio de debida diligencia.

c. Dimensión crítica (sesgos, poder, control metodológico por pares)

Esta dimensión introduce el control metodológico por pares, detección de estereotipos y control del daño diferencial. Merece especial énfasis el control metodológico por pares, entendido como la revisión técnica realizada por otra persona con competencia profesional y científica equivalente, orientada a evaluar la consistencia metodológica del producto, la suficiencia del respaldo, la separación entre hechos e inferencias, la declaración de límites y la identificación de posibles sesgos antes de que el documento produzca efectos jurídicos o institucionales. Cuando un texto pretenda influir de manera relevante en una imputación, un dictamen o una resolución, deberá someterse a un control metodológico orientado a identificar inferencias excedidas, sesgos, omisiones de límites y vacíos relevantes. También es central distinguir forma de método, y en dónde prevalece la ciencia (el fondo) sobre el estilo (la forma). Por otro lado, es vital la vigilancia del sesgo y el daño diferencial. Aquí se revisa si hay estereotipos, variables sustitutas o inferencias sobre poblaciones en situación de vulnerabilidad que funcionen como valoración indirecta; si aparece riesgo de trato desigual, se eleva la exigencia y se aplican salvaguardas reforzadas con perspectiva de interseccionalidad. Finalmente, deben valorarse los costos de contradicción, esto es, se evalúa si la otra parte podría impugnar con acceso razonable a registros; si no, el producto tiende a privatizar la verdad pública.

A partir de lo anterior, y precisamente para que el control humano experto opere como un mecanismo práctico de auditabilidad externa, proponemos una secuencia frugal de trabajo. La secuencia se organiza en pasos simples, pero acumulativos, en los que cada uno deja huellas documentales que facilitan la contradicción real y hacen visible qué se sabe, qué se infiere y qué permanece abierto:

- Planeación: definir con precisión qué se busca probar, qué derechos están en juego y qué rol tendría la IA generativa. Implica traducir la demanda en preguntas resolubles, identificar qué afirmaciones serían "fuertes" (y, por tanto, exigentes de respaldo), y decidir desde el inicio si la IA solo apoyará tareas de ordenamiento/redacción o si se le pedirá producir análisis con potencial de circular a carpeta o dictamen.

- Verificación de insumos: identificar qué existe, qué falta y qué es sensible antes de cualquier interacción con la IA. Aquí se realiza un inventario crítico sobre qué documentos y registros

están disponibles, en qué estado se encuentran, y qué limitaciones presentan. Al mismo tiempo, se marca la información sensible y se define qué puede usarse como insumo y qué debe excluirse o anonimizarse.

- Contrastación: verificar consistencia entre fuentes; marcar contradicciones y evitar que la IA las homologue por estilo. En lugar de buscar "cuadrar la carpeta", el objetivo es mapear convergencias y discordancias. Se identifica qué se confirma entre registros independientes, qué puntos se tensionan y qué contradicciones deben quedar explícitas. Si se usa IA generativa, su papel debe limitarse a organizar comparativamente lo ya documentado (p. ej., matrices de coincidencias y discrepancias) sin reescribir contradicciones.

- Análisis: producir interpretaciones dentro de límites, separar planos y usar la IA solo como soporte de método. La IA puede apoyar en la formulación de preguntas analíticas, en ordenar evidencia o en identificar posibles incoherencias, pero el operador debe mantener el control, justificar inferencias, declarar supuestos y evitar el exceso interpretativo.

- Elaboración: redactar de forma revisable, distinguiendo descripciones, interpretaciones y conclusiones. Se trata de escribir de manera que terceras partes puedan reconstruir el camino seguido, a saber: que cada afirmación fuerte remita a respaldo, que las interpretaciones estén señaladas y que las conclusiones no excedan la base disponible. Si la IA generativa fue utilizada para redactar o reformular, debe quedar claro que su uso es estilístico u organizativo, no probatorio, y que el contenido sustantivo fue verificado contra registros.

- Control metodológico por pares: someter el producto a examen crítico para detectar inferencias excedidas, omisiones de límites, mezclas de planos y efectos de autoridad por estilo o por dispositivo. Antes de circular un texto con peso, se activa revisión interna o externa orientada a preguntas de auditabilidad.

- Control de sesgo y de daño diferencial: evaluar impactos desiguales, variables sustitutas y la impugnabilidad práctica del producto. En este paso se revisa si el texto incorpora estereotipos o inferencias poblacionales que funcionen como imputación indirecta, y si el uso de IA pudo amplificar esos sesgos por la vía de patrones aprendidos o por la forma de la redacción. Además, se valora si la contraparte pudiera impugnar con acceso razonable a registros y a la trazabilidad del proceso.

Lineamientos éticos, salvaguardas de derechos humanos y estándar mínimo de uso responsable

Como se ha sostenido a lo largo del presente artículo, el uso generalizado de IA generativa es una realidad cada vez más presente. Por eso, esta sección pretende sintetizar el contenido del artículo en una serie de recomendaciones y límites para el uso responsable de la herramienta:

a. Recomendaciones para uso de apoyo

En la práctica cotidiana, muchos usos de IA generativa pueden ser compatibles con los derechos humanos si se mantienen en el terreno del apoyo y se conserva la agencia humana. Aquí la herramienta se admite como auxiliar para organizar información, facilitar lectura asistida, preparar preguntas, señalar inconsistencias, ayudar a formular hipótesis alternativas y ordenar la planeación de diligencias.

b. Recomendaciones para usos con pretensión de peso decisivo

El nivel de análisis crítico, la revisión y el control de la IA generativa aumentan cuando el producto pretende influir de manera decisiva (dictamen, imputación o resolución). En otras palabras, el estándar no castiga el uso de la herramienta, solo impide que una prosa fluida se convierta en autoridad sin rastro.

c. Prohibiciones institucionales

Hay escenarios en los que la única forma de proteger los derechos humanos es detener el uso de la IA. Estas prohibiciones pretenderían evitar daño y teatro probatorio en condiciones de crisis. La primera es la imposibilidad de trazabilidad de la información en usos decisorios: en fenómenos de alto impacto, se admite la prohibición cuando no es posible mitigar el riesgo de vulneración de derechos humanos (13). La segunda barrera es prohibir la perfilación, la estigmatización o las inferencias de culpabilidad por variables sustitutas: las correlaciones, el contexto o los rasgos poblacionales no pueden convertirse en "prueba del acto". La tercera es impedir la automatización de decisiones o recomendaciones punitivas: la decisión de imputar, priorizar persecución penal o valorar credibilidad no se delega en sistemas automatizados. Y la cuarta es prohibir la extracción y circulación de datos sensibles sin propósito legítimo y minimización de daño, incluyendo datos biométricos, de salud mental y datos de niñas, niños y adolescentes y otras poblaciones en situación de especial vulnerabilidad.

d. De la crisis al método: correctivos sin burocracia adicional

Un enfoque integrador permite la comprensión investigativa para, así, formular hipótesis, revisar la calidad de la investigación y conectar evidencia conductual, victimología forense y contexto ecológico con enfoque diferencial (8,20,21). En lo práctico, esto se expresa en tres movimientos mínimos. Primero, articulación transversal del caso: romper la departamentalización y forzar la conexión entre las diligencias y quienes las ejecutan e interpretan, para detectar contradicciones y vacíos relevantes (8). Segundo, trabajo activo con hipótesis y refutación, para exigir hipótesis alternativas y discriminadores; enumerar rutas y mapear vacíos, pero no para cerrar investigaciones (20). Tercero, modelo ecológico y vulnerabilidad: describir la vulnerabilidad estructural y el riesgo de sesgo como partes de la debida diligencia, evitando estereotipos y atribuciones ilegítimas (21).

e. Recomendaciones por rol

Para el Estado, se exige separación de planos (no convertir contexto o medios en prueba del acto) y salvaguardas reforzadas ante evidencia sensible, prohibiendo inferencias por variables sustitutas. En

peritajes, el mínimo es tratar la integridad de la evidencia como requisito: declarar condiciones de obtención, alcance y límite; explicitar vacíos y/o ausencias; y activar control metodológico por pares en productos con pretensión decisoria. Para el poder judicial, el estándar se expresa como admisibilidad práctica, es decir, sin respaldo localizable y límites, el texto no puede pesar decisivamente y cuando exista riesgo de daño en derechos humanos, se eleva la exigencia por riesgo-proporcionalidad. Para defensa y asesoría jurídica, el eje es el derecho a la prueba, entendiéndose como el acto de exigir acceso a registros, fuentes y bitácoras; impugnar opacidad; exigir límites y vacíos, y si no existen, pedir que el producto no pese; además, oponerse a estereotipos o uso de variables sustitutas como forma de imputación indirecta.

f. Formación y capacitación para operadores

Finalmente, toda vez que la inmersión de IA generativa en la justicia es un hecho, la respuesta no puede limitarse a "adiestrar en herramientas" (16). En términos mínimos, implica alfabetización en su uso forense (distinguir descripción, interpretación y conclusión; identificar respaldo, límites y vacíos), alfabetización por poder (reconocer cuándo el estilo técnico sustituye al método y cuándo la legibilidad desplaza la contradicción), así como circuitos de control metodológico por pares para productos con pretensión decisoria, priorizando casos con evidencia sensible o alto riesgo. El principio básico del uso de IA debe ser el control humano experto como auditoría permanente y la perspectiva de derechos humanos como brújula ética y metodológica.

Conclusión

En este artículo se ha sostenido que el estándar propuesto incide en tres dimensiones que, en contextos de crisis forense, resultan inseparables: el derecho a la verdad, la preservación de la memoria institucional y las condiciones mínimas para la no repetición. En escenarios donde el Estado pierde capacidad real de investigar con la debida diligencia, el expediente tiende a convertirse en el mundo, en donde lo que está escrito adquiere un peso operativo desproporcionado frente a lo que no se documentó, y el estilo técnico puede funcionar como sustituto del rastro. En ese circuito, la introducción de IA generativa amplifica el riesgo de construcción de narrativas violatorias de derechos fundamentales. Por eso, la sugerencia que atraviesa el texto es que cuando faltan registros esperables, se documentan vacíos y se abren diligencias. En esta clave, la impugnabilidad y la trazabilidad son condiciones de posibilidad del derecho a la prueba, de la defensa adecuada y del derecho a la verdad.

En ese terreno, la IA puede contribuir a la memoria cuando se usa para ordenar, comparar, construir cronologías, localizar contradicciones y señalar qué falta.

En contextos de graves afectaciones, además, la justicia no puede agotarse en la reparación individual. En ese sentido, el estándar mínimo aquí propuesto tiene vocación transformadora por tres vías: desincentiva el cierre por relato al distinguir entre apoyo y decisión, y exigir límites explícitos;

reduce el daño diferencial al activar salvaguardas reforzadas y prohibir inferencias por variables sustitutas; y desprivatiza la verdad pública al exigir auditabilidad y trazabilidad práctica cuando un enunciado pretende peso decisivo.

Se trata, entonces, de pensar con la IA y no delegar en ella el pensamiento, implica sostener el control humano experto como eje central del acto de garantizar el acceso a la justicia. Finalmente, queda enunciar con claridad que este artículo propone mínimos operativos para un uso responsable en contextos forenses, pero no sustituye reglas probatorias ni criterio jurisdiccional. Su implementación, necesariamente, deberá ajustarse a capacidades institucionales reales y a garantías efectivas de acceso a defensa y asesoría, porque sin condiciones materiales de impugnación y sin trazabilidad practicable, no hay auditabilidad externa posible ni acceso al derecho a la verdad.

Contribuciones de autoría (CRediT)

Todas las personas autoras participaron activamente en la conceptualización, metodología, validación, redacción del borrador original, revisión y edición del texto.

Financiamiento

La elaboración del presente manuscrito no contó con financiamiento ni apoyo económico de ninguna institución pública, privada o agencia de financiamiento.

Conflictos de interés

Los autores declaran que no existen conflictos de interés.

Declaración sobre el uso de inteligencia artificial generativa

En la preparación de este manuscrito se utilizó ChatGPT (modelo GPT-5.2 Thinking) como herramienta de apoyo para tareas de redacción y edición (p. ej., mejora de claridad, cohesión y estilo). La herramienta no se empleó para generar contenido empírico original ni para sustituir el juicio académico de las autorías. El texto final, así como la selección y precisión de las citas y referencias, fue revisado y validado por las personas autoras, quienes asumen plena responsabilidad por el contenido publicado. Las ideas y los fundamentos del artículo, así como todo el trabajo intelectual, es responsabilidad exclusiva de las personas autoras, quienes supervisaron el trabajo de redacción con IA generativa.

Referencias

[1] National Institute of Standards and Technology (US). Artificial intelligence risk management framework: Generative artificial intelligence profile. Gaithersburg: National Institute of Standards and Technology; 2024. ↑Ver página 3

[2] Mündler N, He J, Jenko S, Vechev M. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation [Internet]. arXiv [cs.CL]; 2023. Disponible en: http://arxiv.org/abs/2305.15852 ↑Ver página 3

[3] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? En: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency [Internet]. New York: ACM; 2021. Disponible en: http://dx.doi.org/10.1145/3442188.3445922 ↑Ver página 3

[4] Corte Interamericana de Derechos Humanos. Caso González y otras ("Campo Algodonero") vs. México. Excepción Preliminar, Fondo, Reparaciones y Costas. Sentencia de 16 de noviembre de 2009 [Internet]. San José, Costa Rica: Corte IDH; 2009. Disponible en: https://www.corteidh.or.cr/docs/casos/articulos/seriec_205_esp.pdf ↑Ver página 3

[5] Corte Interamericana de Derechos Humanos. Caso Velásquez Rodríguez vs. Honduras. Fondo. Sentencia de 29 de julio de 1988 [Internet]. San José, Costa Rica: Corte IDH; 1988. Disponible en: https://www.corteidh.or.cr/docs/casos/articulos/seriec_04_esp.pdf ↑Ver página 3

[6] Coronado Mares AI. Análisis de la evidencia conductual, victimología y perspectiva de derechos humanos. Identificación Humana [Internet]. 2020; Disponible en: https://www.identificacionhumana.mx/analisis-de-la-evidencia-conductual ↑Ver página 4

[7] Turvey B, Coronado A. Criminal psychology: Forensic examination protocols. México: Forensic Press; 2022. ↑Ver página 4

[8] United Nations. Governing AI for humanity: Final report [Internet]. New York: United Nations; 2024. Disponible en: https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf ↑Ver página 5

[9] Nodal Silva VA. Psicología forense con perspectiva de derechos humanos: Fundamentos teóricos y metodológicos para la investigación psicosocial. Valencia: Tirant Humanidades; 2025. ↑Ver página 6

[10] National Research Council, Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics, Policy and Global Affairs, Committee on Science, Technology, and Law, Committee on Identifying the Needs of the Forensic Sciences Community. Strengthening forensic science in the United States: A path forward. Washington D.C.: National Academies Press; 2009. ↑Ver página 7

[11] President's Council of Advisors on Science and Technology. Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods [Internet]. Washington: Executive Office of the President; 2016. Disponible en: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf ↑Ver página 7

[12] United Nations Population Fund. Elevating Rights and Choices for All: Guidance Note for Applying a Human Rights Based Approach to Programming [Internet]. New York: UNFPA; 2020. Disponible en: https://www.unfpa.org/sites/default/files/pub-pdf/2020_HRBA_guidance.pdf ↑Ver página 7

[13] Office of the United Nations High Commissioner for Human Rights. The Right to Privacy in the Digital Age [Internet]. Geneva: OHCHR; 2020. Disponible en: https://www.ohchr.org/en/documents/thematic-reports/ahrc5117-right-privacy-digital-age ↑Ver página 7

[14] Council of Europe. Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council of Europe Treaty Series No. 225) [Internet]. Strasbourg: Council of Europe; 2024. Disponible en: https://rm.coe.int/1680afae3c ↑Ver página 7

[15] Tabassi E. Artificial Intelligence Risk Management Framework AI RMF (1.0). Gaithersburg: National Institute of Standards and Technology; 2023. https://doi.org/10.6028/NIST.AI.100-1 ↑Ver página 7

[16] Unesco. Recommendation on the Ethics of Artificial Intelligence [Internet]. Paris: Unesco; 2021. Disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381137 ↑Ver página 7 ↑Ver página 12

[17] Fiscalía General de la Nación. Protocolo para la caracterización victimológica en graves afectaciones contra los derechos humanos. Bogotá: Fiscalía General de la Nación; 2022. ↑Ver página 8

[18] Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos. Protocolo de Berkeley sobre las Investigaciones en Fuentes Abiertas Digitales [Internet]. Ginebra: Oacnudh; 2023. Disponible en: https://www.ohchr.org/sites/default/files/2024-01/Berkeley-Protocol-Spanish_0.pdf ↑Ver página 8

[19] Grother P, Ngan M, Hanaoka K. Face Recognition Vendor Test Part 3: Demographic Effects. Gaithersburg: National Institute of Standards and Technology; 2019. https://doi.org/10.6028/NIST.IR.8280 ↑Ver página 8

[20] Turvey BE. Criminal profiling: An Introduction to Behavioral Evidence Analysis. 5th edition. Elsevier; 2025. ↑Ver página 11

[21] UN Women and Office of the United Nations High Commissioner for Human Rights. A Model Protocol for Responding to Violence Against Women and Girls [Internet]. New York: UN Women. 2014. Disponible en: https://www.unwomen.org/en/digital-library/publications/2014/11/model-protocol ↑Ver página 11