Antistio

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

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

Original Article

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

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

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

[1] Doctorate in Psychology, Equipo Independiente de Investigación en Ciencias Forenses y Derechos Humanos, México.
[2] Master's degree in Forensic Sciences, Equipo Independiente de Investigación en Ciencias Forenses y Derechos Humanos, México.
[3] Doctorate in Human Rights, Equipo Independiente de Investigación en Ciencias Forenses y Derechos Humanos, México.

CC BY NC SA

Article history

Date received: 09/04/2026

Date accepted: 27/08/2026

Antistio, vol. 3, 2026, e5897

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.

Correspondence: victor.nodal4177@academicos.udg.mx

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

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.

Introduction

The forensic crisis tends to create a structural human rights problem, in which the State loses its actual capacity to investigate with due diligence and, in that vacuum, the content of investigation files can shift from the verifiable trail toward a compelling but unsubstantiated narrative, that is, toward what "sounds" complete, coherent, and technically well-written, a practice known in criminal jargon as

"making the investigation file add up." In this scenario, artificial intelligence (AI) emerges as a force capable of accelerating two opposing trajectories: on the one hand, it can strengthen the hard work of investigation (organizing scattered information, detecting inconsistencies, mapping evidentiary gaps, harmonizing in accordance with due diligence, and testing competing hypotheses); on the other hand, it can reinforce narrative closure, producing a persuasive text that fills gaps with plausibility without offering sufficient evidence to audit the claims or, in the worst-case scenario, generating narratives based on prejudices, stereotypes, and discriminatory discourse. It follows, therefore, that in forensic contexts, the use of generative AI requires an explicit grounding in human rights to prevent the narrative form from replacing verification, to maintain the practical likelihood of adversarial challenge (enabling traceability, source verification, and effective rebuttal), and to avoid differential harm to populations in vulnerable situations, as well as in cases with greater exposure to biases and inequalities.

For the purposes of this article, "generative artificial intelligence" is understood as the set of systems capable of producing text (and, where applicable, summaries or reformulations) that appears coherent, based on instructions and documentary inputs, as part of a socio-technical process supporting analysis and writing tasks (1). Its critically problematic feature in forensic contexts is not automation itself, but rather its ability to fill epistemic gaps with persuasive narratives, generating plausible outputs that may incorporate unverified or inconsistent content, with the potential to create effects of undue credibility (2,3). Therefore, the analysis is based on forensic crisis scenarios as extreme cases, but the guidelines and safeguards proposed here can be extrapolated to other forensic contexts, adjusting the level of control and traceability according to the risk of harm and the decision-making weight of the generated output, in accordance with the principles of proportionality and human rights-based impact assessment (1).

Furthermore, this article adopts a regional perspective, grounded in the Inter-American Human Rights System (IAHRS). Consequently, the categories of due diligence, judicial guarantees, judicial protection, the right to the truth, a differentiated approach, and the State's duty to investigate are understood in light of the obligations established in the American Convention on Human Rights (ACHR) and their progressive development in the case law of the Inter-American Court of Human Rights (IACtHR) (4,5). Judicial decisions and national provisions may specify or expand these obligations within each legal system, but their systematic examination exceeds the scope of this paper, whose purpose is to propose a technical-scientific standard suitable for regional application and subsequent harmonization with the domestic law of each State. In the absence of a binding inter-American instrument specifically dedicated to the use of artificial intelligence, this proposal develops an applied interpretation of pre-existing treaty obligations in light of the particular risks posed by generative AI, without superseding national rules on evidence, data protection, professional liability, or procedure.

The Risk of Using Generative Artificial Intelligence in Forensic Contexts

In crisis scenarios, criminal investigations tend to become fragmented; as a result, records that would normally be expected are missing, the chain of custody is compromised, and official communications multiply without traceability or follow-up. Amid this deterioration, form can end up supplanting due diligence. The result is forensic theater, that is, the appearance of method without sufficient integration of the forensic trail and the evidence (6).

Institutional legibility, that is, the system's capacity to construct and interpret narratives, becomes especially critical during crises. Institutional systems typically operate through necessary simplifications: they classify, standardize, and aggregate information to make it manageable, in order to construct a case theory that can be presented in court. However, when the environment is degraded, that legibility can take the place of the trail and the evidence, and the document then becomes an "operational fact," even if it is incomplete. In this scenario, AI can accelerate and support the production of texts that are plausible but insufficient in evidentiary and scientific terms, which, in practice, function as facts on paper. The main problem is not "error" in a strictly technical sense, but rather apparent sufficiency, where a plausible text appears that, due to its internal coherence and tone, reads like a conclusion.

From the perspective of access to justice and due process, the consequence is critical: without a competent investigation, it is not possible to clearly establish whether a crime occurred or to determine who the victim is (7).

Based on the foregoing, we identify recurring elements that should be explicitly described, as they function as typical modes of error when generative AI is incorporated without sufficient controls and without a methodological review of the case file:

1. Replacing investigative procedures with summarization: this occurs when the tool is used to "close out" tasks that actually require investigation (verification of sources, cross-checking accounts, identifying gaps, reconstructing sequences, etc.). Synthesis can be useful as a support tool, but it becomes harmful if it replaces basic investigative steps.

2. Authority based on style and blurring of analytical levels: this occurs when the credibility of the content does not stem from a verifiable trail but rather from a technical tone, the fluidity of the argument, or the very fact that it was produced "with AI." Style can create an illusion of rigor, and the tool can function as a seal of accuracy, unwittingly blurring interpretive levels that must remain distinct, for example, facts and hypotheses, description and interpretation, documented data and inference, evidence and explanation.

3. Increased difficulty in refutation: this involves raising the practical difficulty of challenging the claims made. A highly coherent text may require investing more resources to separate data from inferences, trace sources, and dismantle claims, thereby creating asymmetry.

4. Differential harm due to stereotypes or proxy variables: This refers to unequal impacts when the system reproduces stereotypical patterns or uses correlations with sensitive attributes (e.g., geographic location, education level, occupation, language, or criminal history). The error is not always explicit; it is often expressed as implicit attributions, normalized suspicions, or biased interpretations of credibility, intent, or dangerousness.

5. Retrospection and digital open-source evidence: Generative AI systems construct their outputs based on patterns learned from previously curated databases and corpora, so that the coherence

of the result stems not only from the case inputs but also from decisions regarding selection, exclusion, and prioritization made during earlier stages of training or fine-tuning. This curation may incorporate biases, systematic omissions, or representation criteria that pose significant risks to the protection of human rights; thus, without specialized training, a legal professional may not be able to distinguish when the text reflects verifiable evidence from the case file or when it is being shaped by the model in question and the sources to which it has access. In such contexts, AI can reinforce preexisting inequalities: whoever controls resources controls narratives (8).

6. Integration without guidelines: Given the widespread use of the tool, it is easy to infer that the adoption of AI in the field of justice is already occurring on a large scale; however, the availability of institutional guidelines and training often lags far behind its day-to-day use. A human rights-compliant response requires proportionate governance, that is, controls tailored to the risk and the decision-making weight of the output, as well as specialized training, so that the output can be examined, discussed, and challenged with sufficient traceability.

For educational purposes, the following hypothetical example will illustrate these phenomena: a justice system official is investigating a disappearance. The case file contains three interviews with partially contradictory accounts; the phone records analysis is not yet complete; and one investigative step remains to verify the victim's last known location. In order to present a progress report on the investigation, the investigator asks a generative artificial intelligence tool to prepare a report using the available information. The tool produces a clear, well-structured, and seemingly complete document. In it, the contradictions between the interviews disappear because they are integrated into a single coherent narrative; the absence of the phone records analysis is no longer perceived as a significant gap, and the pending investigative step no longer appears as a necessity for continuing the investigation, but rather as a secondary detail. Anyone reading the document might conclude that the investigation is essentially well-established, even though the case file still contains substantial uncertainties. The problem is not that the tool has fabricated facts, but that the coherence of the writing creates an appearance of sufficiency that shifts attention away from what still needs to be investigated toward a narrative that seems to satisfactorily explain what happened. At that point, the summary ceases to be a methodological aid and begins to replace the investigation itself. It is precisely this shift from the verifiable trail toward a narrative of apparent sufficiency that this paper refers to as narrative closure.

Human Rights Perspective: What It Requires and How It Prevents Risks

In an adversarial system, the human rights perspective functions as a cross-cutting standard of operation for any actor involved: it requires effective adversarial challenge, checks on punitive power, due diligence and prevention, and reparations for harm with a differentiated approach. Its purpose is to prevent form from replacing the trail and to ensure that inequality of resources does not ultimately become de facto inequality. In broader terms, this standard is grounded in a transdisciplinary framework that seeks coherence among assumptions of knowledge, understanding of reality, and practical criteria for intervention, so that the methodology retains its meaning for the protection, promotion, and recognition of human dignity (9).

Based on this logic, the approach starts from the operational premise that individuals are rights-holders with legal and social capacity to demand the fulfillment of state obligations. This point is key to the use of AI in forensic contexts, where it is not enough for the output to be "useful" or "efficient"; it must be compatible with obligations of due diligence, checks on power, and procedural safeguards.

In methodological terms, this perspective translates into guiding principles applicable to any practitioner, such as the presumption of innocence, the right to evidence, an adequate defense and effective adversarial challenge, due process, reasonable doubt, elevated standards when the State seeks to restrict rights, the burden of proof resting with the State, and a focus on controlling harm and the abuse of power.

Translated into operational requirements for the use of AI tools, this perspective imposes, at a minimum, several conditions, including (but not limited to) the following:

1. Due diligence: no summary can replace investigative steps; AI, in any case, can only support, organize, or open pathways, but cannot "close" matters on its own.

2. Guarantee of procedural rights and effective adversarial challenge: any statement intended to carry weight must be subject to review; that is, it must be clear what is being asserted, where the supporting material is, what its limitations are, and what information is missing.

3. Differentiated and intersectional approach: When there is a risk of unequal impact, enhanced safeguards must be activated to prevent differential harm and the reproduction of biases, particularly in AI-generated texts and analyses used for interpretation.

4. Right to the truth: in the forensic crisis, narrative closure creates a form of institutional amnesia, because it normalizes the notion that "there is nothing more," even when expected records are missing and plausible alternatives remain unexplored. Public truth depends on a verifiable trail and also on the explicit acknowledgment of what is unknown. Under standards of traceability and human oversight, AI must become a tool for documenting the truth.

In this regard, documenting gaps constitutes an act of institutional honesty and the fulfillment of an obligation that allows for the identification of avenues of investigation, prevents false closures, and preserves the possibility of redress. Hence, the standard must promote three basic practices: 1) preserving and describing the available evidence trail, 2) stating the limits of inference, and 3) documenting relevant gaps that prevent a conclusion.

The same logic applies to the technical minimums related to the integrity of evidence and to negligence. In this context, negligence may consist of failing to comply with standards or protocols, whether by action or omission, directly affecting the possibility of adversarial challenge (7). In terms of evidence, the reliability and scope of forensic methods must be justified and communicated, including

their limitations and, where applicable, potential error rates (10,11). Therefore, when a statement is intended to carry decisive weight, it must be worded in a way that permits review and must answer, with sufficient traceability, basic questions: what is being asserted, where the supporting material is, what its limitations are, what is missing, and what safeguards were applied (12).

From the perspective of inter-American human rights law, the use of AI tools does not alter the regime governing the attribution of state responsibility. When a State decision supported by AI contributes to a flawed investigation, to an unsubstantiated charge, or to an expert report based on unverifiable inferences, the central legal issue is not whether the error was "human" or "technological," but rather whether the State failed in its duty to organize its entire institutional apparatus to ensure an investigation conducted with due diligence and in accordance with judicial guarantees. According to inter-American case law, States are liable for the acts and omissions of their agents, as well as for the absence of sufficient controls to prevent violations of the rights protected by the ACHR. Consequently, the incorporation of generative AI does not diminish the State's duties of supervision, reason-giving, traceability, and expert human oversight; on the contrary, it heightens the requirement to demonstrate that every conclusion with legal effects remained subject to critical review, was open to challenge, and was supported by verifiable evidence.

Expert Human Oversight as a Guarantee of Contestability and External Auditability

In criminal justice, the guiding principle is the risk-proportionality framework; that is, the greater the risk of human rights violations, the greater the need for controls; and when that risk cannot be mitigated, the use of AI in functions carrying decision-making weight must be prohibited (13,14). To make this operational without adding unnecessary bureaucracy, technical risk management frameworks can be employed that prioritize traceability, evaluation, and governance throughout the life cycle of AI systems (15). In the case of generative systems, the risks associated with the production of plausible text, undetected errors, and the degradation of evidence require enhanced controls and explicit documentation of limits (1). As a global framework, UNESCO's recommendation on AI ethics establishes principles of human oversight, transparency, fairness, and accountability that are compatible with this standard (16).

In response to these needs, the proposal calls for review and oversight based on a tiered methodological analysis. The central idea is that not every task requires the same level of control. The intensity of oversight is determined by function (support or decisive weight) and by risk (differential harm, sensitive data, opacity). This framework does not assume that all institutions have the same operational capacity. The intensity of controls must be aligned with the principle of risk proportionality and with available institutional capacities, without compromising the minimum standards compatible with human rights. Thus, for support tasks or low-risk tasks, review by the person responsible for the output may be sufficient; when the result is intended to carry decision-making weight or there is a high risk of infringing on fundamental rights, additional methodological controls must be incorporated in accordance with the institution's capabilities. The standard, therefore, requires ensuring that the level

of supervision is proportional to the legal risk and that no conclusion with significant implications is exempt from expert human oversight.

In this regard, the proposal outlined in the institutional protocol for victimological characterization in serious human rights violations (17) can serve as methodological support because it organizes the work into three dimensions and a sequence of steps aimed at strengthening its evidentiary orientation without replacing investigative steps. This model, applied to the use of AI, focuses on the method and the traceability of the process. This model is presented below:

a. Technical Dimension (Inputs, Safekeeping, Traceability)

The purpose of this dimension is to control what is included, what is excluded, and what is protected to prevent the use of defective inputs (data without a source, without context regarding its collection, or containing unduly exposed sensitive information, etc.). The technical dimension establishes minimum standards, which include defining the scope of the data, recommending that before using generative AI, one define which records are valid for the task (official records, interviews, expert reports, etc.) and prohibit feeding the tool with undifferentiated mixtures of facts, accounts, opinions, and rumors when the output may be included in a case file or expert report. It also requires recording the origin and conditions of collection; that is, for each relevant input, one must note how it was obtained and what limitations it has (e.g., partial copy, secondary source, incomplete transcription, editing). For digital open-source evidence, the threshold is stricter: without a trail of the process, the output cannot carry weight (18). Added to this is the protection of sensitive information, for which enhanced safeguards must be implemented when dealing with children and adolescents, people with disabilities, indigenous communities, victims, biometric data, members of the LGBT+ community, human rights defenders, intimate information, etc., with the guiding principle being to minimize harm by using only what is necessary for the legitimate purpose (19). Finally, practical traceability is required: recording which version of the text was generated, with what inputs, and within what limits.

b. Methodological Dimension (Questions, Hypotheses, Cross-Checking, Limits)

This dimension focuses on the logic used to investigate the case. AI is only acceptable as a support tool if it strengthens the method (formulating questions, organizing evidence, proposing lines of inquiry, and detecting contradictions). Therefore, the first minimum requirement is problem formulation, that is, converting vague or biased requests into answerable questions; for example, instead of asking that "the tool determine," it is more useful to ask what evidence exists, what is missing, and what can be concluded within limits. A second minimum requirement is working with alternative hypotheses. When the text is intended to carry decisive weight, at least two competing hypotheses are required; the tool can help list them, but the user must justify why one holds up better and what missing evidence would alter the conclusion. The third minimum requirement is cross-checking and consistency. Here, the method requires seeking convergence among sources and detecting discrepancies. Added to this are explicit

limitations, such as the requirement that every interpretation must state what cannot be concluded, that is, the natural limits of inference. Finally, relevant gaps must be described by indicating what is missing, why it matters, and what investigative step would resolve it; documenting gaps is an obligation stemming from the principle of due diligence.

c. Critical Dimension (Biases, Power, Methodological Peer Review)

This dimension introduces methodological peer review, the detection of stereotypes, and the control of differential harm. Methodological peer review deserves special emphasis; it is understood as a technical review conducted by another person with equivalent professional and scientific competence, aimed at evaluating the methodological consistency of the document, the sufficiency of the supporting material, the distinction between facts and inferences, the declaration of limits, and the identification of potential biases before the document produces legal or institutional effects. When a text is intended to have a significant influence on a charge, an expert report, or a decision, it must undergo methodological review aimed at identifying overreaching inferences, biases, omissions of limitations, and relevant gaps. It is also essential to distinguish between form and method, and to recognize where science (substance) takes precedence over style (form). Furthermore, vigilance regarding bias and differential harm is vital. Here, we examine whether there are stereotypes, proxy variables, or inferences about populations in vulnerable situations that function as indirect assessments; if there is a risk of unequal treatment, the standards are raised, and enhanced safeguards are applied from an intersectional perspective. Finally, the costs of adversarial challenge must be assessed; that is, we evaluate whether the other party could challenge the findings with reasonable access to records; if not, the output tends to privatize public truth.

Based on the above, and precisely so that expert human oversight functions as a practical mechanism for external auditability, we propose a streamlined workflow. The workflow is organized into simple yet cumulative steps, each of which leaves a documented trail that facilitates effective adversarial challenge and makes visible what is known, what is inferred, and what remains open:

- Planning: precisely define what one seeks to prove, what rights are at stake, and what role generative AI would play. This involves translating the request into answerable questions, identifying which claims would be "strong" (and therefore require substantiation), and deciding from the outset whether the AI will only support organizing and drafting tasks or whether it will be asked to produce analyses that could potentially be included in a case file or expert report.

- Verification of Inputs: Identify what exists, what is missing, and what is sensitive before any interaction with the AI. Here, a critical inventory is conducted to determine which documents and records are available, what condition they are in, and what limitations they present. At the same time, sensitive information is flagged, and it is determined what can be used as input and what must be excluded or anonymized.

- Cross-checking: Verify consistency across sources; flag contradictions and prevent the AI from smoothing them over for stylistic reasons. Rather than trying to "make the investigation file add

up," the goal is to map out points of agreement and disagreement. Identify what is confirmed across independent records, which points are in tension, and which contradictions must be made explicit. If generative AI is used, its role should be limited to comparatively organizing what has already been documented (e.g., matrices of matches and discrepancies) without rewriting contradictions.

- Analysis: Produce interpretations within limits, separate different levels of analysis, and use AI only as a methodological support. AI can assist in formulating analytical questions, organizing evidence, or identifying potential inconsistencies, but the user must maintain control, justify inferences, state assumptions, and avoid excessive interpretation.

- Drafting: Write in a way that allows for review, distinguishing between descriptions, interpretations, and conclusions. The goal is to write in such a way that third parties can reconstruct the path taken, namely: that every strong claim refers to supporting material, that interpretations are clearly marked, and that conclusions do not exceed the available basis. If generative AI was used to draft or rephrase text, it must be clear that its use is stylistic or organizational, not probative, and that the substantive content has been verified against records.

- Methodological peer review: Subject the work to critical examination to detect overreaching inferences, omissions of limits, blurring of levels of analysis, and effects of authority arising from style or the tool. Before circulating a text that carries weight, an internal or external review focused on auditability questions is initiated.

- Bias and differential harm control: assess unequal impacts, proxy variables, and the practical contestability of the output. In this step, the text is reviewed to determine whether it incorporates stereotypes or population-based inferences that function as indirect imputations, and whether the use of AI may have amplified these biases through learned patterns or the wording used. Furthermore, an assessment is made as to whether the other party could challenge the text with reasonable access to records and the traceability of the process.

Ethical Guidelines, Human Rights Safeguards, and a Minimum Standard for Responsible Use

As argued throughout this article, the widespread use of generative AI is an increasingly prevalent reality. Therefore, this section aims to summarize the article's content into a series of recommendations and limits for the responsible use of the tool:

a. Recommendations for Supportive Use

In everyday practice, many uses of generative AI can be compatible with human rights if they remain within the realm of support and human agency is preserved. Here, the tool is accepted as an aid for organizing information, facilitating assisted reading, preparing questions, pointing out inconsistencies, helping to formulate alternative hypotheses, and organizing the planning of investigative steps.

b. Recommendations for Uses Intended to Carry Decisive Weight

The level of critical analysis, review, and oversight of generative AI increases when the output is intended to have a decisive influence (an expert report, a charge, or a decision). In other words, the standard does not penalize the use of the tool; it merely prevents fluid prose from becoming authority without a trail.

c. Institutional Prohibitions

There are scenarios in which the only way to protect human rights is to halt the use of AI. These prohibitions would aim to prevent harm and evidentiary theater in crisis situations. The first is the impossibility of ensuring information traceability in uses intended to carry decision-making weight: in high-impact situations, a prohibition is permissible when it is not possible to mitigate the risk of human rights violations (13). The second barrier is to prohibit profiling, stigmatization, or inferences of guilt based on proxy variables: correlations, context, or population traits cannot become "proof of the act." The third is to prevent the automation of decisions or punitive recommendations: the decision to file charges, prioritize criminal prosecution, or assess credibility is not delegated to automated systems. And the fourth is to prohibit the extraction and circulation of sensitive data without a legitimate purpose and harm minimization, including biometric data, mental health data, and data on children, adolescents, and other populations in situations of particular vulnerability.

d. From Crisis to Method: Corrective Measures Without Additional Bureaucracy

An integrative approach enables investigative understanding, thereby allowing for the formulation of hypotheses, the review of investigative quality, and the connection of behavioral evidence, forensic victimology, and ecological context through a differentiated approach (8,20,21). In practice, this is expressed through three minimal steps. First, cross-cutting coordination of the case: breaking down departmental silos and forcing connections between investigative procedures and those who carry them out and interpret them, in order to detect relevant contradictions and gaps (8). Second, actively working with hypotheses and their refutation to require alternative hypotheses and factors that discriminate between them; identifying pathways and mapping gaps, but not to close investigations (20). Third, the ecological model and vulnerability: describing structural vulnerability and the risk of bias as part of due diligence, avoiding stereotypes and illegitimate attributions (21).

e. Recommendations by Role

For the State, separation of analytical levels is required (not treating context or media coverage as proof of the act), along with enhanced safeguards for sensitive evidence and a prohibition on inferences based on proxy variables. In forensic examinations, the minimum requirement is to treat the integrity of evidence as a prerequisite: stating collection conditions, scope, and limitations; explicitly noting gaps and/or absences; and implementing methodological peer review for outputs intended to carry decision-making weight. For the judiciary, the standard is expressed as practical admissibility; that is, without locatable supporting material and limits, the text cannot carry decisive weight, and when there is a risk

of harm to human rights, the requirement is heightened in proportion to the risk. For the defense and legal counsel, the focus is on the right to evidence, understood as demanding access to records, sources, and logs; challenging opacity; demanding limits and gaps, and, if they do not exist, requesting that the output carry no weight; and opposing stereotypes or the use of proxy variables as a form of indirect imputation.

f. Education and Training for Justice System Practitioners

Finally, since the integration of generative AI into the justice system is now a reality, the response cannot be limited to "training in tools" (16). At a minimum, this entails literacy in its forensic use (distinguishing between description, interpretation, and conclusion; identifying supporting material, limits, and gaps), literacy concerning power (recognizing when technical style supplants method and when legibility displaces adversarial challenge), as well as mechanisms for methodological peer review of outputs intended to carry decision-making weight, prioritizing cases involving sensitive evidence or high risk. The basic principle governing the use of AI must be expert human oversight as an ongoing audit and a human rights perspective as an ethical and methodological compass.

Conclusion

This article has argued that the proposed standard addresses three dimensions that, in contexts of forensic crisis, are inseparable: the right to the truth, the preservation of institutional memory, and the minimum conditions for non-repetition. In scenarios where the State loses its actual capacity to investigate with due diligence, the case file tends to become the world, where what is written takes on disproportionate operational weight compared to what was not documented, and technical style can function as a substitute for the trail. In this cycle, the introduction of generative AI amplifies the risk of constructing narratives that violate fundamental rights. Therefore, the suggestion running throughout the text is that when expected records are missing, gaps should be documented and investigative steps initiated. In this regard, contestability and traceability are conditions that make possible the right to evidence, an adequate defense, and the right to the truth.

In this context, AI can contribute to memory when used to organize, compare, construct timelines, identify contradictions, and highlight what is missing.

Furthermore, in contexts of serious harm, justice cannot be limited to individual redress. In this sense, the minimum standard proposed here has the potential to bring about transformation in three ways: it discourages narrative closure by distinguishing between support and decision-making and requiring explicit limits; it reduces differential harm by activating enhanced safeguards and prohibiting inferences based on proxy variables; and it deprivatizes public truth by requiring auditability and practical traceability when a statement is intended to carry decisive weight.

The point, therefore, is to think with AI rather than delegate thought to it; this entails maintaining expert human oversight as the central element in guaranteeing access to justice. Finally, it must be clearly stated that this article proposes operational minimums for responsible use in forensic contexts, but it does not replace rules of evidence or judicial judgment. Its implementation must necessarily align with actual institutional capacities and effective guarantees of access to defense and legal counsel, because without the material conditions for challenge and without practicable traceability, there can be no external auditability or access to the right to the truth.

Author Contributions (CRediT)

All authors actively participated in the conceptualization, methodology, validation, drafting of the original manuscript, and review and editing of the text.

Funding

The preparation of this manuscript did not receive funding or financial support from any public or private institution or funding agency.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Statement on the Use of Generative Artificial Intelligence

In preparing this manuscript, ChatGPT (model GPT-5.2 Thinking) was used as a support tool for writing and editing tasks (e.g., improving clarity, cohesion, and style). The tool was not used to generate original empirical content or to replace the authors' academic judgment. The final text, as well as the selection and accuracy of citations and references, was reviewed and validated by the authors, who assume full responsibility for the published content. The ideas and foundations of the article, as well as all intellectual work, are the sole responsibility of the authors, who supervised the writing process using generative AI.

References

[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. ↑See page 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. Available at: http://arxiv.org/abs/2305.15852 ↑See page 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. Available at: http://dx.doi.org/10.1145/3442188.3445922 ↑See page 3

[4] Inter-American Court of Human Rights. Case of González et al. ("Cotton Field") v. Mexico. Preliminary Objection, Merits, Reparations and Costs. Judgment of November 16, 2009 [Internet]. San José, Costa Rica: IACtHR; 2009. Available at: https://www.corteidh.or.cr/docs/casos/articulos/seriec_205_esp.pdf ↑See page 3

[5] Inter-American Court of Human Rights. Case of Velásquez Rodríguez v. Honduras. Merits. Judgment of July 29, 1988 [Internet]. San José, Costa Rica: IACtHR; 1988. Available at: https://www.corteidh.or.cr/docs/casos/articulos/seriec_04_esp.pdf ↑See page 3

[6] Coronado Mares AI. Analysis of behavioral evidence, victimology, and the human rights perspective. Human Identification [Internet]. 2020; Available at: https://www.identificacionhumana.mx/analisis-de-la-evidencia-conductual ↑See page 3

[7] Turvey B, Coronado A. Criminal psychology: Forensic examination protocols. México: Forensic Press; 2022. ↑See page 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 ↑See page 5

[9] Nodal Silva VA. Forensic Psychology from a Human Rights Perspective: Theoretical and Methodological Foundations for Psychosocial Research. Valencia: Tirant Humanities; 2025. ↑See page 5

[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. ↑See page 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. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf ↑See page 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. Available at: https://www.unfpa.org/sites/default/files/pub-pdf/2020_HRBA_guidance.pdf ↑See page 7

[13] Office of the United Nations High Commissioner for Human Rights. The Right to Privacy in the Digital Age [Internet]. Geneva: OHCHR; 2020. Available at: https://www.ohchr.org/en/documents/thematic-reports/ahrc5117-right-privacy-digital-age ↑See page 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. Available at: https://rm.coe.int/1680afae3c ↑See page 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 ↑See page 7

[16] Unesco. Recommendation on the Ethics of Artificial Intelligence [Internet]. Paris: Unesco; 2021. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 ↑See page 7 ↑See page 12

[17] Office of the Attorney General. Protocol for Victimological Characterization in Serious Human Rights Violations. Bogotá: Office of the Attorney General; 2022. ↑See page 8

[18] Office of the United Nations High Commissioner for Human Rights. Berkeley Protocol on Digital Open Source Investigations [Internet]. Geneva: OHCHR; 2023. Available at: https://www.ohchr.org/sites/default/files/2024-01/Berkeley-Protocol-Spanish_0.pdf ↑See page 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 ↑See page 8

[20] Turvey BE. Criminal profiling: An Introduction to Behavioral Evidence Analysis. 5th edition. Elsevier; 2025. ↑See page 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. Available at: https://www.unwomen.org/en/digital-library/publications/2014/11/model-protocol ↑See page 11