UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA
T U N J A

TÍTULO MONOGRAFIA

Inteligencia Artificial Generativa: ética, sesgos y sostenibilidad

PROPONENTE(S)
Juan Felipe López González
1.051.064.280
2313543

John Jairo Vergel Camacho
1.002.365.839
2327667

DIRECTOR
Diego Alejandro Vela Beltrán

Tunja
Fecha de presentación 06/05/2026

Tabla de contenidos

RESUMEN

La Inteligencia Artificial Generativa (IAG) constituye una de las transformaciones tecnológicas más profundas y aceleradas de las últimas décadas. Sistemas como GPT-4, Claude, Gemini, DALL-E, Midjourney y Stable Diffusion han demostrado capacidades sin precedentes para generar texto, imágenes, audio, video y código con un nivel de sofisticación que redefine las fronteras entre la producción humana y la automatizada. Esta irrupción tecnológica, sin embargo, no se produce en un vacío social: se despliega sobre estructuras históricas de desigualdad, concentración del poder tecnológico y asimetrías en el acceso a la información que los modelos de IAG absorben, reproducen y, en muchos casos, amplifican a una escala sin precedentes históricos.

La presente investigación constituye una revisión sistemática de literatura científica de alto impacto, ejecutada bajo el protocolo PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), sobre 130 fuentes académicas indexadas publicadas entre 2016 y 2026, seleccionadas a partir de un universo inicial de 104.660 registros distribuidos en cuatro bases de datos internacionales: IEEE Xplore, EBSCO Research, Scopus (Elsevier) y Web of Science (Clarivate). Tras la eliminación de duplicados e irrelevantes se obtuvieron 293 registros únicos que, sometidos a revisión de texto completo bajo criterios estrictos de inclusión y exclusión, conformaron el corpus final de 130 fuentes. La investigación examina la influencia de los sesgos cognitivos, la calidad de los datos de entrenamiento y el impacto ambiental en el desarrollo ético y sostenible de los modelos de IAG, con énfasis sectorial en los ámbitos de la salud y la justicia, y con una proyección específica hacia las implicaciones para el contexto latinoamericano. La elección del protocolo PRISMA 2020 garantizó la transparencia, reproducibilidad y rigor del proceso, documentando con precisión cada decisión de inclusión o exclusión a través de las cuatro fases: identificación, cribado, elegibilidad e inclusión.

Los resultados identificaron y caracterizaron seis categorías principales de sesgo en los sistemas de IAG: sesgos de representación, de medición, de evaluación, de implementación, específicos de los Grandes Modelos de Lenguaje (LLMs) y de carácter histórico-estructural. Estos sesgos no operan de manera independiente sino en interacción dinámica y mutuamente reforzante a lo largo del ciclo de vida completo del modelo. En el ámbito de la salud, la evidencia revisada documenta cómo los modelos de IA entrenados con registros electrónicos tienden a confundir el acceso diferencial a la atención médica con

la necesidad clínica real, perjudicando de manera sistemática a los grupos con menor acceso al sistema de salud. En el ámbito de la justicia, para predecir la probabilidad de reincidencia se revela cómo variables aparentemente neutras pueden actuar como proxies de la raza, reproduciendo discriminación racial bajo una apariencia de objetividad técnica.

Respecto a las estrategias de mitigación evaluadas, la investigación documenta una notable proliferación de propuestas en la literatura: técnicas de preprocesamiento de datos (rebalanceo, generación sintética, cegamiento de atributos sensibles), intervenciones algorítmicas durante el entrenamiento (restricciones de equidad, aprendizaje adversarial), ajustes de postprocesamiento y marcos de gobernanza como el framework NoBIAS y los lineamientos del Reglamento Europeo de Inteligencia Artificial (AI Act). No obstante, la evidencia disponible sobre efectividad es sistemáticamente moderada, contextual y acompañada de compromisos no resueltos: mejorar la equidad intergrupal implica con frecuencia reducciones de hasta diez puntos porcentuales en la exactitud global del modelo, y reducir un tipo de sesgo en LLMs tiende a amplificar otros tipos de manera sistemática (Chand et al., 2026). Esta circularidad evidencia que no existe una solución técnica universal: la elección entre distintas estrategias de mitigación es fundamentalmente una decisión normativa que requiere deliberación pública y participación de las comunidades afectadas.

Las estrategias de IA Verde identificadas poda de modelos, cuantización, destilación del conocimiento y hardware especializado pueden reducir el consumo energético entre un 13% y un 115%, aunque solo el 23% de los estudios sobre Green AI involucran socios industriales, revelando una brecha significativa entre el conocimiento técnico disponible y su adopción práctica. La dimensión ambiental de la IAG representa una urgencia sistémica insuficientemente integrada en los marcos regulatorios y éticos existentes, y su resolución requiere cambios profundos en los incentivos de investigación, en las métricas de éxito de la academia y la industria, y en las políticas energéticas que rigen el funcionamiento de la infraestructura computacional global.

Se concluye que el desarrollo ético y sostenible de la IAG requiere intervenciones simultáneas en múltiples niveles: gobernanza técnica de los datos, marcos regulatorios vinculantes como el AI Act, mecanismos de auditoría independiente, transparencia algorítmica, participación de las comunidades afectadas en el diseño de los sistemas y consideraciones ambientales integradas en el ciclo de vida completo de los modelos. Para

América Latina, y en particular para Colombia, el desafío adicional es superar la dependencia de modelos desarrollados en el Norte Global que no capturan adecuadamente las particularidades lingüísticas, culturales, epidemiológicas y socioeconómicas de las poblaciones locales, configurando lo que la literatura denomina “fallo de transporte algorítmico”. Revertir esta condición de vulnerabilidad tecnológica requiere no solo inversión en infraestructura técnica y formación de talento, sino la construcción de una perspectiva epistémica propia sobre la IA que sitúe a Colombia como sujeto activo del orden tecnológico contemporáneo.

Palabras clave: Inteligencia Artificial Generativa, Sesgos Algorítmicos, Ética en IA, Sostenibilidad Tecnológica, Datos de Entrenamiento, Mitigación de Sesgos.

ABSTRACT

Background: Generative Artificial Intelligence (GAI) represents one of the most profound and accelerated technological transformations in recent decades. Systems such as GPT-4, Claude, Gemini, DALL-E, Midjourney, and Stable Diffusion have demonstrated unprecedented capabilities to redefine the boundaries between human and automated production. However, this disruption unfolds upon historical structures of inequality, power concentration, and informational asymmetries that GAI models absorb, reproduce, and amplify.

Methodology: This research constitutes a systematic literature review of high-impact scientific literature, executed under the PRISMA 2020 protocol. The study analyzes 130 indexed academic sources published between 2016 and 2026, selected from an initial universe of 104,660 records across four international databases: IEEE Xplore, EBSCO Research, Scopus (Elsevier), and Web of Science (Clarivate). The investigation examines the influence of cognitive biases, training data quality, and environmental impact, with a sectoral emphasis on health and justice and a specific projection toward the Latin American context.

Results: The study identified six primary categories of bias: representation, measurement, evaluation, implementation, LLM-specific, and historical-structural. In healthcare, evidence shows that models trained on electronic records often confuse differential access to care

with actual clinical needs. In justice, seemingly neutral variables serve as proxies for race, reproducing discrimination under technical objectivity. Regarding mitigation, while techniques like data rebalancing, adversarial learning, and frameworks (NoBIAS, EU AI Act) exist, their effectiveness is moderate and involves significant trade-offs; improving intergroup equity often results in a decrease of up to 10 percentage points in global accuracy.

Sustainability: "Green AI" strategies—such as model pruning, quantization, and knowledge distillation—can reduce energy consumption by 13% to 115%. However, only 23% of Green AI studies involve industrial partners, revealing a gap between technical knowledge and practical adoption.

Conclusions: Ethical and sustainable GAI development requires simultaneous interventions: binding regulatory frameworks, independent audits, and algorithmic transparency. For Latin America and Colombia, the challenge lies in overcoming "algorithmic transport failure"—the dependence on models from the Global North that fail to capture local linguistic, cultural, and socioeconomic nuances. Reversing this vulnerability requires not only technical investment but the construction of a unique epistemic perspective that positions the region as an active subject in the contemporary technological order.

Keywords: Generative Artificial Intelligence, Algorithmic Bias, AI Ethics, Technological Sustainability, Training Data, Bias Mitigation.

1 INTRODUCCIÓN

La Inteligencia Artificial Generativa (IAG) constituye una de las transformaciones tecnológicas más profundas de las últimas décadas. Sistemas como GPT-4, Claude, Gemini o Midjourney han demostrado capacidades sin precedentes que redefinen las fronteras entre la producción humana y la automatizada. En poco tiempo, han pasado de la investigación académica especializada a permear sectores críticos como la educación, la salud, la justicia, los medios y las finanzas (Martínez-Rolán et al., 2025). Esta rápida irrupción plantea preguntas fundamentales sobre la naturaleza del conocimiento, la autoría,

la responsabilidad y la justicia, convirtiendo a la IAG en un asunto de primer orden en la agenda pública global.

Sin embargo, el entusiasmo legítimo por las capacidades de la IAG no puede ocultar la magnitud de los desafíos que su desarrollo acelerado introduce. El problema central que esta investigación aborda radica en que los sesgos estructurales incorporados durante el proceso de diseño y entrenamiento de los modelos, las deficiencias sistemáticas en la calidad y representatividad de los datos utilizados, y el significativo consumo de recursos energéticos y computacionales asociado a la operación de estos sistemas afectan de manera directa la equidad, confiabilidad y sostenibilidad de la IAG. La literatura científica revisada documenta con solidez empírica que el sesgo en los modelos de IAG no constituye un error técnico aislado sino una expresión estructural de las asimetrías de poder y representación de las sociedades donde estos sistemas son desarrollados (Henao Henao y Hernández Mora, 2025; Pessach y Shmueli, 2022). Un hallazgo central documentado en el corpus el denominado “iceberg del sesgo” (Xiang, 2026), desarrollado en detalle en los resultados de esta investigación demuestra que los modelos alineados para parecer neutrales no eliminan el sesgo sino que lo desplazan hacia capas profundas de representación interna, donde resulta significativamente más difícil de detectar y corregir.

En el ámbito de la salud, los modelos de IA entrenados con registros electrónicos tienden a confundir el acceso diferencial a la atención médica con la necesidad clínica real, perjudicando sistemáticamente a los grupos más vulnerables (Chen et al., 2024). En el ámbito de la justicia, el sistema COMPAS reproduce disparidades raciales históricas bajo una apariencia de objetividad técnica (Alikhademi et al., 2022). Adicionalmente, el entrenamiento de modelos a gran escala genera huellas de carbono considerables: el entrenamiento de GPT-3 consumió 1.287 MWh de electricidad, generó 552 toneladas de CO_2 equivalente y requirió 700.000 litros de agua para refrigeración de servidores (Alzoubi y Mishra, 2024), planteando dilemas de sostenibilidad ambiental de primer orden en un contexto de urgencia climática global.

El panorama regulatorio internacional añade una dimensión adicional de urgencia a este análisis. La adopción del Reglamento Europeo de Inteligencia Artificial (AI Act), aprobado por el Parlamento Europeo en 2024, ha inaugurado una nueva era en la gobernanza de los sistemas de IA, estableciendo obligaciones diferenciadas según el nivel de riesgo de cada aplicación y exigiendo, para los sistemas de alto riesgo, garantías

explícitas de calidad de datos, transparencia, supervisión humana y ausencia de discriminación. Este hito regulatorio, sin precedentes en la escala de sus ambiciones, plantea la pregunta de si los marcos normativos son capaces de seguir el ritmo del desarrollo tecnológico, y qué sucede en los países que aún no cuentan con regulaciones equivalentes, como ocurre en la mayor parte de América Latina y en particular en Colombia. La ausencia de marcos regulatorios nacionales no implica la ausencia de riesgos: la adopción de sistemas de IAG en contextos sin gobernanza adecuada traslada esos riesgos directamente sobre las poblaciones más vulnerables.

La pregunta de investigación que guía el presente estudio es: ¿Cómo influyen los sesgos cognitivos, la calidad de los datos de entrenamiento y el impacto ambiental en el desarrollo ético y sostenible de la Inteligencia Artificial Generativa? Para responder a esta pregunta, se adoptó un enfoque de revisión sistemática de literatura científica bajo el protocolo PRISMA 2020, ejecutado sobre un universo inicial de 104.660 registros distribuidos en cuatro bases de datos internacionales IEEE Xplore, EBSCO Research, Scopus (Elsevier) y Web of Science (Clarivate), a partir del cual, tras la eliminación de duplicados e irrelevantes y la revisión de texto completo bajo criterios estrictos de inclusión y exclusión, se conformó un corpus final de 130 fuentes de alto impacto publicadas entre 2016 y 2026. La investigación se estructura en torno a cinco objetivos específicos que abordan, respectivamente, la influencia de los sesgos cognitivos en el comportamiento de los modelos, el papel de los datos de entrenamiento como fuente de sesgos involuntarios, el análisis de casos documentados en sectores críticos, la evaluación de las estrategias de mitigación disponibles y la cuantificación del impacto ambiental de la IAG. La elección de los sectores de salud y justicia como ejes de análisis sectorial no es arbitraria: son los ámbitos donde los errores algorítmicos tienen las consecuencias más directas e irreversibles sobre las personas diagnósticos erróneos, privación de libertad injustificada, estigmatización y donde la necesidad de marcos éticos robustos resulta más imperiosa. La presente monografía aspira a contribuir a ese debate desde una perspectiva latinoamericana, aportando un marco de referencia documentado, riguroso y accesible para quienes deben tomar decisiones sobre el desarrollo, la adopción y la regulación de la IAG en nuestro contexto regional.

Objetivo General

Examinar la influencia de los sesgos cognitivos, la calidad de los datos de entrenamiento y el impacto ambiental en el desarrollo de modelos de Inteligencia Artificial Generativa, mediante una revisión sistemática de literatura científica de alto impacto, con el propósito de valorar sus implicaciones éticas y las estrategias vigentes para su mitigación.

Objetivos Específicos

1. Explorar cómo los sesgos cognitivos influyen en el desarrollo y comportamiento de modelos de IAG
2. Investigar el papel de los datos de entrenamiento en la introducción de sesgos involuntarios
3. Analizar casos documentados donde modelos de IA han mostrado sesgos significativos
4. Evaluar estrategias actuales para mitigar sesgos en modelos de IAG
5. Determinar el impacto ambiental asociado al entrenamiento y operación de modelos de IAG

JUSTIFICACIÓN

La justificación de la presente investigación se articula en cinco dimensiones complementarias e interdependientes: la relevancia social y ética del problema, la urgencia regulatoria del momento histórico, las brechas identificadas en la producción académica existente, la pertinencia específica para el contexto latinoamericano y colombiano, y la contribución metodológica que aporta la integración sistemática de tres ejes temáticos que la literatura existente ha abordado de manera fragmentada. Estas dimensiones no son independientes entre sí: se refuerzan mutuamente y configuran, en conjunto, la razón de ser de un estudio que no pretende simplemente añadir un título más a una bibliografía en expansión acelerada, sino producir conocimiento articulado y accesible que sirva como referente para quienes enfrentan decisiones concretas sobre el desarrollo, la adopción y la regulación de la IAG.

Desde el punto de vista social y ético, el despliegue masivo de sistemas de IAG en decisiones que afectan directamente la vida de las personas el acceso a la atención médica, las sentencias judiciales, las oportunidades de empleo, la calificación crediticia o la evaluación educativa convierte las preguntas sobre equidad y justicia algorítmica en una urgencia de derechos fundamentales. La evidencia revisada en esta investigación

documenta con contundencia que los sesgos en los sistemas de IAG no son errores aleatorios corregibles con ajustes técnicos menores, sino expresiones estructurales de las desigualdades históricas que caracterizan las sociedades donde estos modelos son desarrollados. El análisis sistemático del corpus permitió identificar seis categorías de sesgo de representación, de medición, de evaluación, de implementación, específicos de LLMs e histórico-estructurales que no operan de manera independiente sino en interacción dinámica y mutuamente reforzante a lo largo del ciclo de vida completo del modelo. Se ha evidenciado que los modelos alineados para parecer neutrales no eliminan el sesgo, sino que lo desplazan hacia capas profundas de representación interna donde resulta más difícil de detectar y corregir, haciendo que las auditorías superficiales de neutralidad sean fundamentalmente insuficientes para garantizar la equidad real de un sistema. El caso del sistema COMPAS, que sobreestima sistemáticamente el riesgo de reincidencia para acusados afroamericanos con una tasa de falsos positivos que duplica la de los acusados blancos, y los hallazgos sobre sesgos de género en los modelos de lenguaje que asocian consistentemente roles profesionales de alta jerarquía con el género masculino mientras atribuyen roles de cuidado y trabajo doméstico al género femenino (Muñoz-García et al., 2025), ilustran de manera concreta el potencial de la IAG para amplificar, institucionalizar y legitimar discriminaciones que la sociedad ha identificado como inaceptables. A esto se suma el hallazgo de Mouri Zadeh Khaki y Choi (2026) que demuestra cómo los agentes negociadores autónomos basados en LLMs ofrecen condiciones económicas sistemáticamente menos favorables a compradores latinos que a compradores blancos con una diferencia de margen de USD 5,75 frente a USD 4,72, evidenciando que la discriminación algorítmica opera incluso en ausencia de términos discriminatorios explícitos. La reproducción algorítmica de desigualdades históricas no es un problema abstracto ni futuro: está ocurriendo ahora mismo, en sistemas que toman decisiones que afectan a millones de personas a escala global y, progresivamente, en Colombia y América Latina.

Desde el punto de vista de la urgencia regulatoria, el momento histórico en que se desarrolla esta investigación es singularmente relevante. La aprobación del AI Act europeo en 2024 el primer marco regulatorio vinculante de alcance global sobre inteligencia artificial marca el inicio de una era de gobernanza normativa que plantea desafíos inmediatos tanto para los sistemas ya desplegados como para los que están en desarrollo. Este reglamento

establece obligaciones diferenciadas según el nivel de riesgo de cada aplicación: para los sistemas de alto riesgo utilizados en salud, justicia, educación y servicios esenciales, exige garantías explícitas de calidad de datos, transparencia, supervisión humana y ausencia de discriminación. Su artículo 10(5) va más allá al permitir el tratamiento excepcional de categorías especiales de datos personales incluyendo datos raciales y de salud cuando sea necesario para detectar y corregir sesgos, sentando las bases de lo que Capellà i Ricart (2026) interpreta como una medida de acción positiva algorítmica sin precedente normativo. Esta transición regulatoria genera una demanda real y urgente de conocimiento sistematizado: qué riesgos de la IAG están documentados con solidez empírica, qué estrategias de mitigación han demostrado efectividad real y cuáles solo mejoran métricas superficiales mientras desplazan el problema hacia capas más profundas, como evidencia el concepto del iceberg del sesgo, y qué condiciones institucionales y técnicas favorecen una gobernanza efectiva. La presente investigación contribuye directamente a satisfacer esa demanda al organizar y sintetizar la mejor evidencia disponible bajo un protocolo de revisión sistemática que garantiza la transparencia y reproducibilidad del análisis. En el contexto latinoamericano, donde la Política Nacional de Inteligencia Artificial de Colombia (CONPES 3975, 2019) y su Marco Ético (2021) son documentos de orientación voluntaria sin mecanismos de verificación ni sanciones, este conocimiento sistematizado adquiere un valor estratégico inmediato: provee la base empírica que hace posible argumentar, con evidencia, por qué Colombia necesita marcos normativos vinculantes y cuáles son los umbrales mínimos que esos marcos deben garantizar. La gobernanza blanda, como señalan Papagiannidis et al. (2025), es insuficiente incluso en entornos con alta capacidad institucional como el europeo: si las prácticas de IA responsable frecuentemente no se implementan de manera sistemática cuando son voluntarias, cabe esperar que sus limitaciones sean aún mayores en el contexto colombiano, donde las capacidades de supervisión técnica están en una fase inicial de desarrollo.

Desde el punto de vista de las brechas en la producción académica en sesgos de LLMs los estudios se centran exclusivamente en el idioma inglés y una buena parte de los artículos revisados involucran financiación de grandes corporaciones tecnológicas, lo que evidencia que el conocimiento producido sobre IA responsable está en sí mismo sesgado hacia un contexto cultural, lingüístico e institucional particular generando puntos ciegos sistemáticos sobre la experiencia de comunidades no anglófonas y que parte significativa

de ese conocimiento está financiada por los mismos actores cuyas prácticas se pretende regular: el propio campo de estudio del sesgo algorítmico padece el sesgo que pretende diagnosticar. A esta concentración geográfica se suma una concentración temática: el 82% de los estudios sobre sesgo de género en LLMs lo tratan como una variable binaria (masculino/femenino), excluyendo identidades no binarias, transgénero e intersex, lo que significa que los modelos de mitigación desarrollados a partir de esa evidencia pueden no funcionar para las poblaciones excluidas de los estudios de evaluación. La fragmentación disciplinaria añade una capa adicional de problema: los estudios sobre sesgos algorítmicos, los marcos éticos y regulatorios, y la investigación sobre sostenibilidad ambiental de la IA se publican en revistas y conferencias distintas, con lenguajes técnicos diferentes y sin marcos de referencia compartidos, lo que dificulta la comprensión holística de un fenómeno cuyas dimensiones son inseparables. La presente investigación, desarrollada desde una perspectiva latinoamericana y en lengua española, contribuye a contrarrestar el sesgo de producción académica al incorporar explícitamente la dimensión regional y al integrar simultáneamente las tres dimensiones principales sesgos algorítmicos, ética y gobernanza, y sostenibilidad ambiental en un único análisis comprehensivo respaldado por un corpus de 130 fuentes de alto impacto. Esta integración no es meramente aditiva: la revisión revela cómo las tres dimensiones se condicionan mutuamente de maneras que los estudios especializados por separado no pueden capturar, como la relación entre el paradigma de escalamiento sin restricciones, la huella ambiental y la proliferación de sesgos sistémicos.

Desde el punto de vista de la pertinencia latinoamericana y colombiana, el fenómeno denominado en la literatura como “fallo de transporte algorítmico” el deterioro del rendimiento de un modelo cuando se aplica en un contexto geográfico o demográfico diferente al de su entrenamiento original constituye un riesgo crítico y concreto para la adopción de tecnologías de IAG en Colombia y en la región (Calahorrano Latorre, 2025). Los modelos desarrollados en el Norte Global están optimizados para poblaciones anglófonas, con características genéticas, epidemiológicas, lingüísticas y socioculturales específicas, y su extrapolación a contextos latinoamericanos sin validación local puede generar errores sistemáticos con consecuencias graves en sectores como la salud o la justicia. En salud, los modelos predictivos entrenados con registros clínicos norteamericanos o europeos confunden el menor acceso de ciertas poblaciones a la atención médica con una menor necesidad clínica real (Chen et al., 2024), un error que en

el sistema de salud colombiano con sus profundas desigualdades de acceso entre zonas urbanas y rurales, entre regiones y entre grupos étnicos se traduce directamente en discriminación asistencial. En justicia, la adopción de sistemas predictivos en un país con la historia de criminalización diferencial como Colombia, que afecta desproporcionadamente a jóvenes de periferia urbana, líderes sociales y comunidades en territorios de conflicto reproduciría bajo apariencia de objetividad técnica exactamente las mismas discriminaciones que el sistema ha perpetuado históricamente por otros medios. En el caso de los modelos de lenguaje, la dimensión lingüística del fallo de transporte es especialmente aguda: los modelos entrenados mayoritariamente en inglés y español neutro estándar presentan rendimientos inferiores en las variedades dialectales colombianas. Esta exclusión no afecta solo el rendimiento técnico: cuando los sistemas de IA son incapaces de procesar o generar texto en lenguas indígenas, se producen dinámicas donde las instituciones del Estado adoptan herramientas que son literalmente incapaces de servir a sectores significativos de la población que tienen derechos reconocidos a la atención en su lengua. El aporte académico de esta investigación radica, por tanto, en producir un marco de referencia documentado, riguroso y accesible para desarrolladores, reguladores, tomadores de decisiones, comunidades académicas y ciudadanía en general, que les permita comprender los riesgos reales de la IAG, evaluar críticamente las soluciones disponibles y participar informadamente en los debates sobre gobernanza tecnológica que definirán el tipo de sociedad que queremos construir con y a través de la inteligencia artificial. La ausencia de Colombia del debate global sobre IA ética y sostenible es, en última instancia, una forma de delegación política y epistémica: al no participar activamente en la producción del conocimiento y de los marcos normativos que gobiernan el desarrollo de la IA, Colombia delega en otros actores las decisiones sobre qué valores se codifican en las tecnologías que ella adopta y qué definiciones de equidad y sostenibilidad prevalecerán.

Desde el punto de vista de la contribución metodológica y la urgencia ambiental, esta quinta dimensión justifica el diseño de la investigación como revisión sistemática bajo el protocolo PRISMA 2020 por razones que van más allá de la convención académica. La fragmentación disciplinaria descrita en la dimensión anterior no es solo un problema bibliográfico: genera una comprensión incompleta de fenómenos que son, en esencia, interdependientes. El entrenamiento de la inteligencia artificial consume grandes cantidades de electricidad, debido a esto se genera contaminación ambiental a gran escala,

lo cual no es un dato independiente de los sesgos de ese modelo: ambos son consecuencias del mismo paradigma de escalamiento sin restricciones que Schwartz et al. (2020) denominan “IA Roja”, donde el único criterio de éxito es el rendimiento técnico medido en benchmarks, sin consideración del costo ambiental ni del impacto social diferencial sobre distintos grupos. De manera análoga, en la mitigación de sesgos en LLMs” la evidencia empírica de que reducir un tipo de sesgo tiende a amplificar otros (Chand et al., 2026) conecta directamente con la dimensión de gobernanza: si no existe una solución técnica universal, la elección entre estrategias de mitigación es una decisión normativa que exige marcos institucionales capaces de articular valores, deliberar con las comunidades afectadas y rendir cuentas, no solo laboratorios capaces de optimizar funciones de pérdida. La integración sistemática de estas tres dimensiones bajo un corpus unificado de 130 fuentes de alto impacto partiendo de 104.660 registros en cuatro bases de datos internacionales y aplicando criterios rigurosos de inclusión y exclusión bajo el protocolo PRISMA 2020 produce un marco de análisis cuya coherencia interna y cuya capacidad para identificar tensiones transversales no habría sido posible desde ninguna perspectiva disciplinaria aislada. La constatación de que solo el 23% de los estudios sobre IA Verde involucran socios industriales (Verdecchia et al., 2023) a pesar de que las estrategias de eficiencia energética disponibles pueden reducir el consumo entre un 13% y un 115% revela una brecha de implementación que solo es visible cuando la dimensión ambiental se analiza en el mismo marco que los incentivos académicos e industriales. Esta integración constituye, en sí misma, una contribución original al campo cuya ausencia en la literatura existente fue constatada durante el proceso de identificación bibliográfica, y que justifica plenamente el diseño metodológico adoptado.

Alcances y Limitaciones

La presente investigación tiene un alcance temático, sectorial, metodológico y geográfico claramente delimitado, y reconoce un conjunto de limitaciones inherentes a su diseño que deben contextualizarse adecuadamente para interpretar correctamente sus conclusiones.

- *Alcances*

En cuanto al alcance temático, el estudio examina tres ejes interrelacionados: (1) los sesgos algorítmicos en sistemas de IAG, caracterizados en seis categorías de representación, de medición, de evaluación, de implementación, específicos de LLMs e

histórico-estructurales que operan de manera interdependiente a lo largo del ciclo de vida completo del modelo; (2) los marcos éticos, regulatorios y de gobernanza que orientan el desarrollo responsable de estos sistemas, incluyendo el EU AI Act, el framework NoBIAS y la arquitectura de IA Confiable; y (3) el impacto ambiental asociado al entrenamiento e inferencia de modelos a gran escala, abordado desde el paradigma de la IA Verde y sus estrategias de eficiencia energética. La integración de estos tres ejes en un único análisis responde a la evidencia de que operan de manera interdependiente: los sesgos son consecuencia parcial de los incentivos del paradigma de escalamiento que también genera la huella ambiental, y ninguno de los dos puede abordarse sin los marcos de gobernanza que articulan la dimensión normativa.

En cuanto al alcance sectorial, la investigación concentra su análisis aplicado en los sectores de salud y justicia, identificados como los ámbitos donde los errores algorítmicos tienen consecuencias más directas e irreversibles sobre los derechos de las personas: diagnósticos erróneos, privación de libertad injustificada, perpetuación de discriminaciones raciales y socioeconómicas bajo apariencia de objetividad técnica. La elección de estos dos sectores responde también a la disponibilidad de evidencia empírica rigurosa: el corpus identificado concentra el 29% de sus fuentes (38 de 130) en casos sectoriales de salud y justicia. Se abordan de manera secundaria casos documentados en educación, agricultura y medios de comunicación, que aportan perspectivas comparativas sobre los patrones transversales de los sesgos y permiten establecer la generalidad de los hallazgos más allá de los sectores primarios.

En cuanto al alcance metodológico, la investigación constituye una revisión sistemática de literatura ejecutada bajo el protocolo PRISMA 2020, sobre un universo inicial de 104.660 registros distribuidos en cuatro bases de datos internacionales IEEE Xplore, EBSCO Research, Scopus (Elsevier) y Web of Science (Clarivate), de los cuales se obtuvo un corpus final de 130 fuentes de alto impacto artículos publicados en revistas Q1/Q2, actas de conferencias de primer nivel (ACM FAccT, IEEE, NeurIPS), libros académicos y documentos institucionales de referencia publicadas entre 2016 y 2026 en inglés y español. El 75% del corpus corresponde al período 2023-2026, garantizando la máxima actualidad en un campo de rápida evolución; de ese porcentaje, aproximadamente el 52% del total del corpus pertenece específicamente al período 2024-2026, reflejando la explosión investigativa posterior a la masificación de herramientas como ChatGPT, Gemini y Claude.

El alcance analítico es descriptivo-analítico: descriptivo porque caracteriza el estado actual del conocimiento sobre los tres ejes; analítico porque descompone los fenómenos en sus dimensiones constitutivas, examina las relaciones entre ellos y evalúa críticamente las estrategias de mitigación y gobernanza disponibles. La investigación no pretende ser predictiva ni prescriptiva en sentido absoluto, aunque sus conclusiones ofrecen orientaciones fundamentadas para la práctica y la política pública.

En términos de alcance, aunque el corpus abarca la producción global de las bases de datos analizadas, el estudio orienta sus hallazgos específicamente hacia los contextos latinoamericano y colombiano. Esta focalización se materializa en el análisis del fallo de transporte algorítmico, la identificación de poblaciones excluidas de los beneficios de los LLM, el examen del vacío normativo local frente al AI Act europeo y la evaluación de la viabilidad de estrategias como el aprendizaje federado en instituciones del país. Lejos de ser un añadido secundario, la dimensión regional constituye un eje analítico fundamental que vertebra tanto la interpretación de los datos como las conclusiones finales.

- *Limitaciones*

La presente investigación reconoce cuatro limitaciones inherentes a su diseño que condicionan la interpretación de sus conclusiones: la ausencia de experimentación empírica propia, el sesgo lingüístico y geográfico del corpus hacia el Norte Global, la velocidad de evolución del campo que puede haber superado algunos hallazgos, y la exclusión de literatura gris técnica de alta relevancia. Estas limitaciones se describen y contextualizan en detalle en el capítulo de Diseño Metodológico (sección Limitaciones Metodológicas), que es el lugar apropiado para su tratamiento, junto con las estrategias adoptadas para mitigarlas dentro del protocolo PRISMA 2020 aplicado.

2 MARCO DE REFERENCIA

2.1 Antecedentes de la Investigación

El estudio de las implicaciones éticas y sociales de la Inteligencia Artificial (IA) ha experimentado una transformación profunda en la última década, transitando de preocupaciones teóricas sobre la autonomía de las máquinas hacia problemáticas concretas de justicia, equidad, calidad de datos y sostenibilidad. Los antecedentes de esta investigación se remontan a los trabajos fundacionales que identificaron, por primera vez,

la capacidad de los algoritmos para perpetuar y amplificar desigualdades estructurales, sentando las bases de lo que hoy se denomina IA Responsable.

Uno de los hitos más relevantes en este campo fue el trabajo pionero sobre sesgos en sistemas de reconocimiento facial, donde se demostró que dichos sistemas presentaban tasas de error significativamente más elevadas en grupos demográficos minoritarios. Este hallazgo evidenció lo que se denominó la "codificación" de prejuicios sociales en la infraestructura tecnológica y marcó un punto de inflexión en la disciplina, estableciendo la necesidad de auditar los sistemas de IA no únicamente por su eficiencia técnica, sino también por su impacto social diferenciado sobre distintos grupos poblacionales.

Con la irrupción de los Grandes Modelos de Lenguaje (LLM) y la Inteligencia Artificial Generativa (IAG), la discusión académica se complejizó considerablemente. Bender et al. (2021) cuyo artículo fundacional “On the Dangers of Stochastic Parrots” no formó parte del corpus final de esta revisión por criterios de período temporal, aunque su influencia en el campo es ampliamente reconocida alertaron sobre los riesgos inherentes a los modelos de lenguaje a gran escala, no solo desde una perspectiva ética, sino también ambiental, introduciendo el debate sobre la sostenibilidad en la agenda de la IA. Sus contribuciones pusieron de manifiesto que los problemas de estos modelos no son simplemente técnicos, sino que se entrelazan con cuestiones de poder, representación y responsabilidad colectiva. En años más recientes, Martínez-Rolán et al. (2025) y Henao Henao y Hernández Mora (2025) han analizado cómo la IAG reconfigura los escenarios sociales y éticos, señalando que la capacidad de estos modelos para generar contenido altamente realista introduce riesgos inéditos de desinformación, manipulación y erosión de la confianza social a una escala sin precedentes históricos.

Paralelamente, el ámbito de la calidad de los datos ha emergido como un factor determinante en la confiabilidad de los sistemas de IA. De Andrade et al. (2026) documentaron las profundas consecuencias que los problemas de calidad en los registros electrónicos de salud tienen para las aplicaciones de IA en medicina crítica, revelando que la imprecisión, incompletitud e inconsistencia de los datos subyacentes compromete fundamentalmente la validez de los modelos construidos sobre ellos. Esta línea de investigación evidencia que la promesa de la IA en entornos de alto riesgo depende, antes que de la sofisticación algorítmica, de la integridad de los datos que la alimentan.

En el plano regulatorio, los antecedentes son igualmente significativos. La adopción del Reglamento Europeo de Inteligencia Artificial (AI Act) ha inaugurado una nueva era en la gobernanza de estos sistemas, estableciendo obligaciones específicas según el nivel de riesgo de las aplicaciones. Este marco normativo encuentra sus raíces académicas en trabajos como los de Díaz-Rodríguez et al. (2023), quienes propusieron conectar los principios éticos con los requisitos técnicos a través de una regulación responsable. En el ámbito de la salud y la justicia, los antecedentes son particularmente críticos: Chen et al. (2024) documentaron sistemáticamente cómo los modelos basados en registros electrónicos heredan sesgos históricos resultando en diagnósticos menos precisos para poblaciones subrepresentadas, mientras que los sistemas de predicción de reincidencia en el sector judicial han sido ampliamente cuestionados por reproducir disparidades raciales preexistentes. Esta trayectoria de investigación establece el cimiento sobre el cual el presente estudio analiza la intersección entre sesgos, ética, calidad de datos, gobernanza y sostenibilidad en la IAG.

2.2 Bases Teóricas

Para abordar la complejidad de la Inteligencia Artificial Generativa, este marco teórico se articula en ocho pilares fundamentales que estructuran la investigación: (1) la ética, la justicia algorítmica y la gobernanza como fundamento normativo; (2) la tipología y mitigación de sesgos en modelos generativos; (3) los sesgos específicos en los Grandes Modelos de Lenguaje (LLMs); (4) la calidad de los datos de entrenamiento como factor determinante de la equidad; (5) la sostenibilidad tecnológica bajo el paradigma de la IA Verde; (6) las implicaciones éticas y los sesgos en el sector salud; (7) la justicia algorítmica y la policía predictiva; y (8) las estrategias transversales de mitigación y responsabilidad sectorial. Esta estructura permite analizar el fenómeno desde lo conceptual hacia lo aplicado, identificando tanto los desafíos técnicos como las responsabilidades sociales, regulatorias y ambientales que emergen del despliegue masivo de la IAG.

2.2.1 Fundamentos Éticos, Justicia Algorítmica y Gobernanza de la IA

La implementación de sistemas de IA en decisiones que afectan directamente la vida humana desde la evaluación crediticia hasta el diagnóstico médico y las sentencias judiciales exige marcos éticos que superen la mera optimización matemática. La ética en IA ha transitado de principios abstractos hacia marcos operativos de "IA Confiable"

(Trustworthy AI), proceso que ha involucrado a la academia, reguladores, organismos internacionales e industria, y cuyo recorrido conceptual es indispensable para comprender el estado actual del debate.

Teorías de justicia y el diseño algorítmico equitativo

La teoría de la justicia de John Rawls (1971) proporciona una lente teórica especialmente potente para evaluar la equidad de los sistemas algorítmicos. Mediante el concepto del "velo de ignorancia", Rawls propone que las reglas de una sociedad deberían diseñarse sin que sus autores conozcan de antemano la posición que ocuparán en ella. Aplicado al diseño de IAG, este principio implica que los desarrolladores no deberían construir modelos que beneficien sistemáticamente a grupos privilegiados hablantes de inglés, hombres, poblaciones de determinada etnia a expensas de otros. Si los diseñadores de modelos estuvieran detrás de un velo de ignorancia, probablemente priorizarían la equidad y la robustez para los casos marginales, ya que podrían pertenecer a cualquiera de esos grupos poblacionales.

Sin embargo, como señalan Pessach y Shmueli (2022) en su revisión exhaustiva sobre equidad en aprendizaje automático, traducir conceptos filosóficos de justicia en métricas matemáticas operativas es un desafío de primer orden. Existe una tensión estructural entre distintas definiciones de equidad paridad demográfica, igualdad de oportunidades, calibración, donde satisfacer una implica con frecuencia violar otra, lo que hace necesario que los diseñadores de sistemas tomen decisiones normativas explícitas sobre qué tipo de equidad se prioriza en cada contexto de uso. Decker et al. (2025) amplían esta perspectiva argumentando que la justicia procedimental la equidad en los procesos de toma de decisiones algorítmicas, no solo en sus resultados requiere mecanismos de participación pública y rendición de cuentas. Su investigación demuestra que cuando los ciudadanos tienen la oportunidad de comprender y cuestionar las decisiones algorítmicas, la aceptación social y la legitimidad del sistema aumentan significativamente. Wenzelburger et al. (2024) complementan esta visión subrayando que el contexto de implementación de los algoritmos sus reglas, incentivos y entorno institucional es tan determinante para su impacto como el diseño técnico mismo.

De la ética principista a la IA Confiable

En respuesta a estas tensiones, la literatura reciente ha convergido en el concepto de IA Confiable como marco integrador. Li et al. (2023) proponen un modelo que identifica ocho dimensiones críticas de confiabilidad: robustez, generalización, explicabilidad, transparencia, reproducibilidad, privacidad, responsabilidad y equidad. Para la IAG, esta distinción es especialmente relevante porque un modelo puede ser técnicamente robusto resistente a ataques adversariales y, al mismo tiempo, profundamente inequitativo en sus salidas, generando contenido sesgado o discriminatorio de manera sistemática.

La explicabilidad resulta especialmente problemática en la IAG. Las redes neuronales profundas que sustentan herramientas como GPT-4 o DALL-E operan como cajas negras, lo que dificulta identificar las causas de una salida discriminatoria: ¿reside en los datos de entrenamiento, en el diseño del algoritmo o en el prompt del usuario? Kaur et al. (2022) argumentan que la confianza en la IA no puede construirse sin transparencia, mientras que Díaz-Rodríguez et al. (2023) proponen un enfoque de "puntos de conexión" que vincula cada principio ético equidad, privacidad, responsabilidad con requisitos técnicos verificables y marcos regulatorios específicos. La IA confiable, desde esta perspectiva, no es un estado estático sino un proceso continuo de alineación entre valores humanos y comportamiento algorítmico.

Gobernanza responsable: marcos normativos y regulatorios

La gobernanza de la IA ha evolucionado desde recomendaciones voluntarias hacia marcos regulatorios vinculantes. Papagiannidis et al. (2025) ofrecen una revisión comprehensiva de la gobernanza de la IA responsable, identificando que los enfoques más efectivos combinan mecanismos formales de regulación con normas de autorregulación sectorial y mecanismos de participación ciudadana. Su análisis revela que la gobernanza efectiva requiere no solo reglas y sanciones, sino también capacidades institucionales, cultura organizacional y recursos para la implementación.

González-Fernández et al. (2025) evidencian que, aunque existe un consenso creciente sobre los principios deseables transparencia, no maleficencia, responsabilidad, privacidad y equidad, la brecha entre los marcos normativos y las prácticas institucionales reales continúa siendo considerable, en parte por la falta de herramientas operativas y en parte por la insuficiente formación de los actores responsables de la implementación. En el

contexto europeo, CASANOVA (2025) analiza los sesgos y la discriminación desde la perspectiva del Derecho comunitario, señalando que el AI Act establece obligaciones específicas para sistemas de alto riesgo: gestión de datos de alta calidad, documentación técnica, transparencia hacia los usuarios, supervisión humana y robustez. CAPELLÀ I RICART (2026) aborda además el tratamiento de categorías especiales de datos personales datos raciales, de salud, religiosos para la detección y corrección de sesgos, argumentando que en ciertos contextos el acceso a estos datos puede ser no solo permisible sino necesario como medida de acción positiva para garantizar la equidad del sistema.

Hu et al. (2025) añaden una dimensión adicional al señalar que la IAG introduce tensiones inéditas en torno a la autoría, la originalidad, la verificabilidad y la responsabilidad editorial en la publicación académica, que los marcos tradicionales de integridad científica no estaban diseñados para abordar. Este es un ejemplo paradigmático de cómo la velocidad del desarrollo tecnológico supera la capacidad adaptativa de los marcos normativos existentes.

2.2.2 Sesgos Algorítmicos y Estrategias de Equidad en IAG

El sesgo en la Inteligencia Artificial Generativa no constituye un fallo técnico aislado, sino una característica sistémica que refleja las asimetrías de poder y representación presentes en la sociedad. La literatura especializada distingue tres categorías de sesgo según su origen en el ciclo de vida del modelo, y la comprensión de esta tipología es condición necesaria para el diseño de estrategias de mitigación efectivas.

Tipología de los sesgos

Los sesgos de representación o de datos emergen cuando los conjuntos de entrenamiento no reflejan adecuadamente la diversidad de la población. En el contexto de los LLMs, esto se manifiesta de forma especialmente pronunciada: dado que la mayor parte de los textos disponibles en internet son producidos por hombres, en inglés, desde perspectivas culturales occidentales y sobre temáticas que interesan a las clases medias con acceso a tecnología, los modelos entrenados sobre estos corpus inevitablemente internalizan estas asimetrías. Li y Deng (2022) demostraron empíricamente que los datasets de reconocimiento de expresiones faciales presentan un sesgo cultural marcado,

evidenciando que los modelos generativos de imágenes entrenados con datos similares reproducen estas distorsiones con alta fidelidad.

Los sesgos algorítmicos o de medición surgen cuando las métricas de optimización favorecen implícitamente a ciertos grupos. Abdel-Aziz et al. (2025) demostraron que las métricas de evaluación convencionales pueden ocultar disparidades sistemáticas en el rendimiento del modelo según el tipo de texto o el dominio cultural de origen; incluso modelos técnicamente sofisticados pueden presentar sesgos sutiles que solo se detectan mediante evaluaciones desagregadas por subgrupos. En el ámbito clínico, Chen et al. (2024) identificaron que los modelos predictivos basados en historiales médicos confunden la accesibilidad al sistema de salud con la necesidad médica real, subestimando la gravedad de las condiciones en poblaciones con menor acceso a la atención sanitaria.

Los sesgos de evaluación se producen cuando los conjuntos de prueba no son representativos de la diversidad poblacional. Glocker et al. (2023) evidenciaron este fenómeno en modelos de detección de enfermedades a partir de radiografías de tórax, encontrando que la codificación algorítmica de características protegidas sexo o raza generaba un rendimiento diagnóstico dispar entre subgrupos. Paradigmáticamente, el modelo no utilizaba estas variables como predictores explícitos, sino que las inferían a partir de patrones sutiles en las imágenes, lo que hace el sesgo difícil de detectar con métodos de auditoría convencionales. Parra et al. (2022) añaden una perspectiva organizacional, documentando que la percepción de sesgos raciales y de género en las recomendaciones de IA influye significativamente en la disposición de los usuarios a cuestionar dichas recomendaciones, generando dinámicas donde los sesgos del sistema interactúan con los sesgos cognitivos de los usuarios de maneras no lineales.

Estrategias de mitigación y sus limitaciones

La mitigación de sesgos en IAG requiere intervenciones en múltiples niveles del ciclo de vida del modelo. Pessach y Shmueli (2022) categorizan las estrategias en tres momentos: el pre-procesamiento (rebalanceo de clases y augmentación de datos antes del entrenamiento), el en-procesamiento (modificaciones algorítmicas que penalizan la dependencia del modelo respecto a atributos sensibles) y el post-procesamiento (ajustes en las salidas para asegurar resultados equitativos). Skaiky et al. (2025) advierten, no obstante, que no existe una solución universal. En la IAG, donde la salida es creativa y

probabilística no una clasificación binaria, la definición de un "resultado equitativo" de referencia es en sí misma un desafío normativo.

Rajawat et al. (2024) proponen un enfoque de auditoría continua que combina métodos cuantitativos métricas de equidad computables con métodos cualitativos revisión experta y participación de comunidades afectadas, subrayando que la identificación de sesgos es tan importante como su mitigación: un sesgo no detectado no puede ser corregido. Alvarez et al. (2024) enfatizan que la solución técnica debe ir acompañada de políticas organizacionales: la gestión de sesgos no es exclusivamente un problema de ciencia de datos, sino un desafío de gobernanza que requiere auditorías independientes, documentación transparente mediante model cards y datasheets, y mecanismos de rendición de cuentas a lo largo de toda la cadena de desarrollo y despliegue.

2.2.3 Sesgos en Grandes Modelos de Lenguaje (LLMs): Tipología y Mitigación

Los Grandes Modelos de Lenguaje GPT-4, Claude, Gemini o Llama constituyen la arquitectura central de la revolución de la IAG. A diferencia de los sistemas discriminativos tradicionales, los LLMs presentan un desafío singular: su naturaleza probabilística y su entrenamiento sobre vastos corpus de texto no solo replican estereotipos, sino que pueden amplificarlos o generar nuevas formas de discriminación mediante el fenómeno de las "alucinaciones". Henao Henao y Hernández Mora (2025) señalan que los LLMs representan un salto cualitativo en las implicaciones éticas de la IA, pues su capacidad para generar texto coherente, persuasivo y contextualmente apropiado los convierte en herramientas potencialmente poderosas tanto para la comunicación humana como para la manipulación y la desinformación.

El problema de la definición: equidad en modelos generativos

Un punto de partida ineludible es la ausencia de consenso sobre qué significa "equidad" en el contexto de los LLMs. Yin et al. (2026) demuestran que las definiciones operativas válidas para modelos clasificadores como la paridad demográfica no son directamente trasladables a modelos generativos, donde la equidad no es una salida binaria sino una distribución de probabilidades sobre millones de tokens. Esta complejidad se agrava por lo que el estudio Bias Research for LLMs is Biased (2026) denomina "sesgo de investigación": la propia literatura tiende a focalizarse en el idioma inglés, en contextos occidentales y en ejes de discriminación visibles género y raza, dejando subrepresentados

los sesgos cognitivos, culturales y lingüísticos del Sur Global. Este meta-problema implica que las soluciones propuestas actualmente pueden ser insuficientes para contextos globales o para grupos demográficos subrepresentados en la academia.

Manifestaciones específicas del sesgo en LLMs

La investigación reciente ha permitido clasificar los sesgos en los LLMs en cuatro categorías interrelacionadas. En cuanto al sesgo de género y representación laboral, Muñoz-García et al. (2025) desarrollaron metodologías para cuantificar el sesgo de género utilizando el corpus BiasBloom, demostrando que los modelos no solo asocian estereotipos, sino que producen texto de menor calidad gramatical cuando se refieren a grupos minoritarios, penalizando lingüísticamente a los mismos. Nomelini y Marcolin (2024) evidenciaron cómo los LLMs, al generar ofertas de empleo, asocian sistemáticamente los cargos de liderazgo con características masculinas y los roles de cuidado con características femeninas, perpetuando barreras estructurales de acceso al mercado laboral.

Respecto al sesgo político e ideológico, Choudhary (2025) realizó un análisis comparativo empírico entre los modelos líderes del mercado GPT-4, Gemini, Perplexity y Claude, encontrando inclinaciones ideológicas sistemáticas y diferenciadas según el proceso de ajuste fino y retroalimentación humana (RLHF) de cada modelo. Rettenberger et al. (2025) advierten que la aparente neutralidad de un modelo suele enmascarar una posición centrista-liberal específica, lo que puede generar desconfianza en usuarios con otras afiliaciones políticas y plantea interrogantes sobre la legitimidad democrática de estos sistemas.

En lo relativo al sesgo territorial y cultural, Sarrionandia et al. (2025) identificaron una "brecha de representación geográfica" en cinco modelos generativos: los sistemas comprenden mejor la realidad del Norte Global y fallan significativamente al contextualizar eventos de España y Latinoamérica en comparación con eventos de Estados Unidos o Reino Unido. Este hallazgo es crítico para la región iberoamericana, donde la adopción de IAG puede profundizar asimetrías epistemológicas ya existentes.

Finalmente, en cuanto a sesgos cognitivos e interaccionales, Lou y Sun (2026) demostraron experimentalmente la presencia del "sesgo de anclaje" en LLMs: los modelos

tienden a depender excesivamente de la información inicial proporcionada en el prompt y a validar premisas falsas del usuario en lugar de corregirlas. Esto tiene implicaciones directas en contextos profesionales de alto riesgo: si un médico o un juez introduce una hipótesis sesgada en su consulta, el modelo tiende a confirmarla y expandirla algorítmicamente.

El dilema de la mitigación

La mitigación de sesgos en LLMs enfrenta una restricción técnica fundamental. Chand et al. (2026), en "No Free Lunch in LLM Bias Mitigation", prueban que las intervenciones para reducir un tipo de sesgo amplifican con frecuencia otros o degradan la capacidad de razonamiento del modelo. Por ejemplo, forzar a un modelo a ser políticamente neutral puede reducir su capacidad para responder preguntas fácticas sobre eventos sociales controvertidos. Ante este desafío, la literatura propone estrategias innovadoras: el uso de prompts metacognitivos que invitan al modelo a identificar sus propias inconsistencias antes de generar la respuesta final (Hills, 2026); la integración de bases de conocimiento externas y verificables para corregir alucinaciones (Yang et al., 2026); y el desarrollo de enfoques específicos por dominio en sectores sensibles como los cuidados de larga duración (Rickman, 2025). Estas propuestas reconocen implícitamente que no existe una solución universal al problema del sesgo en LLMs: la mitigación eficaz exige una combinación de intervenciones técnicas, decisiones normativas y supervisión humana continua.

2.2.4 Calidad de los Datos de Entrenamiento en Sistemas de IAG

La calidad de los datos de entrenamiento constituye uno de los factores más determinantes y frecuentemente subestimados en el comportamiento ético y técnico de los modelos de IAG. El principio de "basura entra, basura sale" (garbage in, garbage out), ampliamente conocido en la informática clásica, adquiere dimensiones particularmente críticas en el contexto del aprendizaje profundo, donde los modelos tienen la capacidad de aprender y amplificar patrones erróneos de manera no lineal y a escala masiva. Comprender las dimensiones de la calidad de los datos es, por tanto, un requisito previo para entender el origen estructural de muchos de los sesgos descritos en las secciones anteriores.

Dimensiones de la calidad de datos

La calidad de los datos puede analizarse desde múltiples dimensiones interrelacionadas. La precisión hace referencia a que los datos reflejen correctamente la realidad que pretenden describir; la completitud indica que no existan valores faltantes sistemáticos que introduzcan sesgos de omisión; la consistencia implica coherencia a lo largo del tiempo y entre distintas fuentes; la actualidad señala que los datos sean suficientemente recientes; y la representatividad asegura que los datos cubran adecuadamente la diversidad de la población objetivo. Cuando alguna de estas dimensiones falla, los modelos construidos sobre esos datos heredan y pueden amplificar las deficiencias de forma exponencial.

De Andrade et al. (2026) realizaron una revisión sistemática sobre los problemas de calidad de datos en registros electrónicos de salud, identificando consecuencias profundas para las aplicaciones de IA en medicina crítica. Sus hallazgos revelan que problemas como la fragmentación de historiales clínicos, los errores de codificación diagnóstica, la variabilidad en las convenciones de registro entre instituciones y los sesgos de selección que determinan qué pacientes y qué información se registran se traducen directamente en modelos de IA con limitaciones sistemáticas que pueden comprometer la seguridad del paciente. Esta investigación subraya que la calidad de los datos no es un problema técnico menor, sino un factor que puede determinar la viabilidad misma de las aplicaciones de IA en entornos clínicos de alta criticidad.

Vajpayee y Khobragade (2024) abordan específicamente el problema del sesgo de datos en la IA sanitaria, señalando que los conjuntos de datos clínicos frecuentemente sobre-representan a pacientes de poblaciones mayoritarias hombres, blancos, de clases medias con acceso regular a atención médica mientras sub-representan a mujeres, minorías étnicas, personas mayores, pacientes con comorbilidades múltiples y poblaciones rurales. Este desequilibrio no es aleatorio, sino que refleja desigualdades estructurales en el acceso a la atención médica que luego se reproducen y potencialmente amplifican en los modelos de IA entrenados con esos datos.

Estrategias para mejorar la calidad de los datos

Las estrategias para mejorar la calidad de los datos de entrenamiento pueden clasificarse en tres niveles. A nivel técnico, incluyen la auditoría y limpieza de datos antes

del entrenamiento, la augmentación de datos para grupos subrepresentados mediante técnicas generativas o de sobremuestreo, el uso de técnicas de aprendizaje federado que permiten entrenar modelos sobre datos distribuidos sin centralizarlos, y la implementación de mecanismos de privacidad diferencial que protejan la información sensible sin sacrificar la representatividad del conjunto de entrenamiento.

A nivel institucional, se requieren políticas de gobierno de datos que establezcan estándares de calidad y responsabilidades claras. Chen et al. (2024) documentaron que en el sector salud, la adopción de estándares interoperables para el intercambio de información clínica como HL7 FHIR puede mejorar significativamente la consistencia y completitud de los datos disponibles para el entrenamiento de modelos de IA. Sin embargo, advierten que la interoperabilidad técnica no garantiza automáticamente la equidad representativa; se requieren esfuerzos adicionales para asegurar que los datos provengan de poblaciones diversas. A nivel regulatorio, marcos como el RGPD europeo y el AI Act establecen requisitos específicos sobre la calidad de los datos en sistemas de alto riesgo, incluyendo que sean relevantes, representativos y apropiados para el propósito del sistema.

El problema de los datos sintéticos

Una alternativa que ha ganado considerable atención es el uso de datos sintéticos generados por modelos de IAG para aumentar o reemplazar datos reales escasos o sesgados. Sin embargo, esta estrategia introduce sus propias tensiones: los datos sintéticos solo pueden ser tan representativos y equitativos como los modelos que los generan, los cuales a su vez dependen de los datos reales con los que fueron entrenados. Este riesgo de circularidad donde los sesgos de los datos reales se reproducen en los datos sintéticos, que luego se utilizan para entrenar nuevos modelos requiere mecanismos cuidadosos de validación y auditoría para evitar la amplificación progresiva de los sesgos originales. Este fenómeno constituye uno de los desafíos más complejos y menos resueltos en la literatura sobre calidad de datos para IAG.

2.2.5 Sostenibilidad Tecnológica: El Paradigma de la IA Verde

Durante décadas, la investigación en IA priorizó la precisión de los modelos sin considerar los costos computacionales y ambientales asociados. Este enfoque, denominado "IA Roja" (Red AI) por Schwartz et al. (2020), está siendo cuestionado de manera creciente por el emergente paradigma de la "IA Verde" (Green AI), que propone

integrar la eficiencia energética y la sostenibilidad ambiental como criterios de evaluación de igual importancia que las métricas de rendimiento técnico. Para la IAG, esta cuestión es especialmente crítica dado el extraordinario costo computacional asociado al entrenamiento e inferencia de modelos a gran escala.

Huella de carbono y consumo energético de los modelos de IAG

Schwartz et al. (2020) fueron pioneros en establecer la distinción entre eficiencia y precisión, argumentando que la comunidad académica debería publicar métricas de eficiencia costo energético, emisiones de CO_2 con la misma prominencia que las métricas de rendimiento. Entrenar un modelo como GPT-3 consume una cantidad de energía equivalente a la de varios hogares durante años, generando una huella de carbono que rara vez se reporta en los artículos que celebran sus logros técnicos.

Verdecchia et al. (2023), en su revisión sistemática sobre IA Verde, sistematizan los métodos para reducir el impacto ambiental: uso de hardware especializado TPUs que puede reducir el consumo energético hasta un orden de magnitud respecto a las GPU convencionales; optimización de hiperparámetros mediante búsqueda eficiente; técnicas de "pruning" que eliminan parámetros del modelo con menor impacto en el rendimiento; y técnicas de destilación del conocimiento, donde un modelo grande transfiere su conocimiento a un modelo más pequeño y eficiente para su despliegue. Bolón-Canedo et al. (2024) amplían el concepto de sostenibilidad al ciclo de vida completo del modelo: en la IAG, la fase de inferencia el uso diario por millones de usuarios puede llegar a superar el consumo energético del entrenamiento inicial, lo que implica que las estrategias de optimización deben extenderse también a la arquitectura de despliegue. Budennyy et al. (2022) proponen herramientas como eco2AI para rastrear las emisiones de carbono en tiempo real, facilitando una "contabilidad del carbono" alineada con los Objetivos de Desarrollo Sostenible (ODS) de la ONU, especialmente el ODS 13 sobre Acción por el Clima.

La IA como solución y como problema ambiental

Un aspecto frecuentemente soslayado es la paradoja del impacto ambiental de la IA: por un lado, es consumidora significativa de energía; por otro, puede ser herramienta poderosa para abordar el cambio climático. Wang et al. (2024) analizaron el impacto de la IA sobre las huellas ecológicas y las transiciones energéticas, concluyendo que el potencial

descarbonizador de la IA solo se realizará si el propio desarrollo de la IA se vuelve sostenible. Yigitcanlar et al. (2021) propusieron el concepto de "IA Verde para ciudades inteligentes", argumentando que la eficiencia tecnológica debe considerarse siempre en relación con el bienestar humano y la justicia social: un modelo eficiente que perpetúa sesgos no puede considerarse verdaderamente verde desde una perspectiva de sostenibilidad integral.

Alzoubi y Mishra (2024) identifican tanto el potencial transformador como los desafíos de implementación de la IA verde: entre los potenciales destacan la optimización de redes eléctricas, la mejora de la eficiencia energética en edificios y el diseño de materiales para la captura de carbono; entre los desafíos, la necesidad de una transición energética en los centros de datos que depende de políticas públicas más que de innovación algorítmica. Gaur et al. (2023), desde la Teoría de Sistemas de Sistemas, concluyen que la IA para la sostenibilidad requiere cambios en los incentivos de investigación y en las métricas de éxito de la academia y la industria. Hossain et al. (2024) y Moghimi et al. (2024) documentan aplicaciones concretas en movilidad urbana y gestión energética en edificios, mientras que Mana et al. (2024) y Melak et al. (2024) amplían esta perspectiva al sector agrícola y ganadero, donde la IA puede contribuir a reducir el uso de insumos y minimizar las emisiones de gases de efecto invernadero.

Xiang et al. (2023) modelizaron el consumo energético de los centros de datos y propusieron estrategias de optimización que pueden reducirlo entre un 20 y un 40%, señalando que la arquitectura del centro de datos es tan importante para la huella de carbono de la IA como la eficiencia de los propios algoritmos. Estas investigaciones colectivamente dibujan un campo en el que la sostenibilidad ambiental no puede separarse de la sostenibilidad social y ética: ambas dimensiones son condición necesaria para un desarrollo genuinamente responsable de la IAG.

2.2.6 Implicaciones Éticas y Sesgos en el Sector Salud

La implementación de IAG en el ámbito sanitario constituye uno de los desafíos éticos más complejos de la actualidad. A diferencia de otros sectores, en salud los errores algorítmicos no se traducen en meras molestias digitales, sino en daños físicos concretos: diagnósticos erróneos, tratamientos inadecuados o la negación de atención a quienes más

la necesitan. La evidencia reciente (2023-2026) confirma que la IAG está reconfigurando la práctica clínica al tiempo que hereda y amplifica desigualdades estructurales preexistentes.

Equidad algorítmica y sesgo en datos clínicos

El problema fundacional de la IA sanitaria radica en la calidad y representatividad de los datos de entrenamiento. Chen, R. J. et al. (2023), en su revisión seminal para Nature Biomedical Engineering, establecen que la equidad algorítmica no es un atributo por defecto, sino un objetivo de diseño que requiere intervenciones activas. Los modelos tienden a aprender correlaciones espurias: un algoritmo puede aprender que ciertos grupos demográficos acceden menos a la atención médica y confundir este determinante social con una menor necesidad clínica, subestimando así la gravedad de sus condiciones. Abràmoff et al. (2023) argumentan que, sin una auditoría de sesgos explícita, los algoritmos exacerban disparidades raciales y socioeconómicas.

Agarwal et al. (2023) proponen marcos de "conciencia de sesgo" (bias-aware) que interconectan el sesgo técnico con la política sanitaria, subrayando que la mitigación debe permear las decisiones de adopción tecnológica y no limitarse al nivel del código. Vajpayee y Khobragade (2024) documentaron que el sesgo de datos en la IA sanitaria tiene consecuencias especialmente graves en oncología, cardiología y enfermedades crónicas donde la detección temprana es crítica para el pronóstico, señalando que los ensayos clínicos han excluido históricamente a mujeres, personas mayores y minorías étnicas, comprometiendo así la base de evidencia sobre la que se construyen los modelos. En el contexto latinoamericano, Calahorrano Latorre (2025) advierte sobre el "fallo de transporte algorítmico": los modelos desarrollados en el Norte Global a menudo fallan al extrapolarse a poblaciones latinoamericanas debido a diferencias genéticas, epidemiológicas y de registro de datos.

Limitaciones de la IAG en predicción clínica

La evidencia empírica reciente urge a la cautela frente al entusiasmo por la IAG clínica. Brown et al. (2025) publican en JAMIA hallazgos que desafían la narrativa de la IAG como solución omnicomprensiva: los LLMs son significativamente menos efectivos que modelos locales especializados en tareas de predicción clínica. Su incapacidad para razonar causalmente sobre fisiopatología limitándose a patrones estadísticos de lenguaje genera riesgos reales de "alucinaciones diagnósticas". En radiología, Gichoya et al. (2023)

documentaron errores sistemáticos: los algoritmos de visión por computadora fallan de forma desproporcionada en imágenes de pacientes de piel oscura o con características anatómicas atípicas. Goetz et al. (2024) señalan que la generalización es el desafío clave para una IA responsable en aplicaciones clínicas: un modelo de alto rendimiento entrenado en un hospital universitario puede fracasar en un centro rural, ampliando la brecha sanitaria en lugar de reducirla.

La crisis de sobreconfianza y la ética clínica

Más allá de la precisión técnica, emerge un problema ético de naturaleza psicológica. Berkowitz et al. (2026) introducen el concepto de "crisis de sobreconfianza": el riesgo real en la IA clínica no radica exclusivamente en la imprecisión de la máquina, sino en la confianza excesiva del médico en la recomendación algorítmica. Cuando un sistema de IAG genera un diagnóstico con tono de seguridad lingüística aunque sea incorrecto, los clínicos tienden a reducir su escrutinio crítico, comprometiendo la supervisión humana necesaria. Goktas y Grzybowski (2025) subrayan que la ética clínica tradicional basada en la autonomía del paciente y la beneficencia del médico se ve perturbada cuando un agente no humano influye de manera determinante en la decisión terapéutica.

Para mitigar estos riesgos, Stanley et al. (2026) proponen validar la equidad algorítmica no solo mediante métricas estadísticas, sino a través de comités éticos multidisciplinarios. Ueda et al. (2024) recomiendan que los sistemas reporten obligatoriamente su nivel de incertidumbre al médico, evitando la "ilusión de competencia". Desde una perspectiva organizacional, Muhammad Kakale y Pinelli (2026) advierten que la presión por la eficiencia económica puede llevar a adoptar sistemas de IAG éticamente inmaduros, priorizando la velocidad sobre la seguridad del paciente. Melo (2024) corrobora esta preocupación en el contexto hispanohablante, señalando que la aplicación de IA en pacientes requiere un consentimiento informado que explique con claridad el rol del algoritmo, algo que raramente ocurre en la práctica clínica actual. Finalmente, Ueda et al. (2024) señalan la paradoja ambiental del sector: el uso de IA para mejorar diagnósticos contribuye a la crisis climática, la cual, a su vez, deteriora los determinantes sociales y ambientales de la salud, reforzando la necesidad de un enfoque integrado donde la sostenibilidad ambiental sea también un pilar de la ética médica.

2.2.7 Justicia Algorítmica y Policía Predictiva

Si en salud el riesgo central de la IA es la negligencia asistida, en el sector justicia y seguridad el riesgo adquiere una gravedad institucional superior: la violación de derechos fundamentales y la perpetuación algorítmica de la discriminación sistémica. La aplicación de algoritmos en la predicción delictiva policía predictiva y en la toma de decisiones judiciales es el campo donde el debate sobre justicia algorítmica alcanza su mayor intensidad y donde las consecuencias de los errores privación de libertad, estigmatización, discriminación racial resultan irreversibles para los individuos afectados.

Fundamentos y riesgos estructurales de la policía predictiva

La policía predictiva utiliza datos históricos de crímenes para anticipar dónde ocurrirán futuros delitos o identificar quién tiene mayor probabilidad de cometerlos. Berk (2021), en su revisión para Annual Review of Criminology, advierte que estos sistemas no predicen el crimen per se, sino las instancias reportadas de crimen. Dado que el reporte policial está sesgado por prácticas de vigilancia excesiva en barrios minoritarios, el algoritmo aprende a concentrar la presencia policial en esos mismos barrios, generando un bucle de retroalimentación que confirma y amplifica su propio sesgo de forma sistemática.

El caso paradigmático es el de Correctional Offender Management Profiling for Alternative Sanctions, en el cual investigaciones periodísticas y académicas revelaron que el sistema sobreestimaba sistemáticamente el riesgo de reincidencia para acusados afroamericanos e infraestimaba dicho riesgo para acusados blancos, a pesar de que sus desarrolladores afirmaban que no utilizaba la raza como variable predictora. Este caso ilustra la diferencia crucial entre sesgo explícito el uso directo de variables sensibles como la raza y sesgo proxy el uso de variables aparentemente neutras, como el código postal o el historial de arrestos, que están correlacionadas con la raza debido a desigualdades estructurales previas. CASANOVA (2025) analiza este fenómeno desde el Derecho antidiscriminatorio europeo, concluyendo que los marcos regulatorios deben evolucionar desde la prohibición de la discriminación directa hacia la prevención activa de la discriminación indirecta generada por proxies algorítmicos.

Alikhademi et al. (2022), en una revisión fundamental publicada en Artificial Intelligence and Law, concluyen que, sin mecanismos de corrección explícitos, la policía predictiva codifica la discriminación racial en la lógica operativa de las fuerzas de seguridad. Kaufmann et al. (2019) añaden que la delegación de la discrecionalidad policial en cajas

negras algorítmicas dificulta el escrutinio democrático de las decisiones. Meijer y Wessels (2019) concluyen que, si bien la eficiencia puede aumentar, el costo en términos de libertades civiles es elevado; Meijer et al. (2021) profundizan en cómo la algoritmización de la burocracia erosiona la capacidad de contextualización humana en decisiones que requieren matiz y consideración de circunstancias individuales.

Sesgos de implementación y el problema de la discrecionalidad

Un aspecto crítico, frecuentemente soslayado en el debate técnico, es la interacción entre los humanos y estas herramientas. Selten et al. (2023) descubrieron que los funcionarios públicos tienden a confiar en las recomendaciones de la IA solo cuando confirman su juicio preexistente. Esto revela que la IA, lejos de corregir los sesgos humanos, los valida y los blinda de crítica: si un oficial tiene un prejuicio contra un grupo y el algoritmo lo marca como "de alto riesgo", la sospecha queda algorítmicamente respaldada. Wenzelburger et al. (2024) destacan que en jurisdicciones donde los jueces utilizan las recomendaciones algorítmicas como punto de partida ejerciendo revisión crítica, los sesgos pueden ser moderados; sin embargo, cuando los algoritmos se convierten en la norma de referencia y los operadores desarrollan una dependencia excesiva "automatización sesgada", los sesgos se amplifican y se legitiman institucionalmente.

Linera (2023), analizando el sistema VioGén para la prevención de la violencia de género en España, muestra cómo la automatización de la evaluación de riesgo puede generar una falsa sensación de seguridad institucional, transfiriendo responsabilidades humanas a un sistema que no puede asumir las consecuencias de sus errores sobre víctimas vulnerables. Asaro (2019) propone, desde una ética del cuidado, reemplazar la concepción del ciudadano como amenaza potencial que subyace a la justicia predictiva por un enfoque que presuma la inocencia y priorice los derechos civiles. Nyawambi y Muchiri (2024) insisten en que la transparencia de los datos de entrenamiento es el primer paso obligatorio para cualquier estrategia de mitigación de sesgo racial en sistemas de justicia algorítmica.

2.2.8 Estrategias de Mitigación y Responsabilidad Sectorial

El panorama descrito en los apartados anteriores exige no solo el diagnóstico de los problemas, sino la construcción de marcos de respuesta robustos y adaptados a las particularidades de cada sector. La literatura reciente converge en una conclusión central:

la mitigación de los sesgos en IAG no puede ser únicamente técnica; requiere una combinación de estrategias de datos, intervenciones algorítmicas y marcos de gobernanza institucional que involucren a todos los actores de la cadena de valor tecnológica.

Estrategias técnicas y de proceso en salud

Hasanzadeh et al. (2025), en npj Digital Medicine, proponen una taxonomía de intervención que abarca desde la limpieza de datos antes del entrenamiento hasta el monitoreo continuo del modelo en producción. La clave, según estos autores, es concebir la mitigación como un proceso iterativo y no como un ajuste estático: los sesgos pueden surgir o evolucionar con el tiempo a medida que cambian las poblaciones, las prácticas clínicas o los patrones de uso del sistema. Chin et al. (2023), en JAMA Network Open, recomiendan que los desarrolladores de IAG colaboren activamente con sociólogos y expertos en salud pública para definir qué significa "justo" en cada contexto clínico, dado que la definición matemática de equidad varía según el escenario en oncología, la penalización por falsos negativos difiere radicalmente de la tolerable en un cribado rutinario.

Polevikov (2023) sugiere que la validación de modelos debe realizarse en múltiples sitios clínicos antes de su despliegue masivo, garantizando que el rendimiento sea equitativo a través de contextos institucionales diversos. Nouis et al. (2025) documentan que los médicos demandan explicabilidad real: necesitan comprender el porqué de una recomendación para incorporarla con juicio crítico, algo que los modelos generativos actuales no siempre pueden ofrecer de manera verificable. Esta demanda de explicabilidad no es solo una preferencia profesional, sino un requisito ético para preservar la autonomía clínica y la responsabilidad médica frente al paciente.

Gobernanza y políticas públicas en justicia

En el ámbito de la justicia, la mitigación requiere menos código y más gobernanza. Siala y Wang (2022) proponen que la evaluación de impacto social sea un requisito obligatorio previo a la implementación de cualquier sistema predictivo, estableciendo una lógica de precaución regulatoria que invierta la carga de la prueba: los desarrolladores deben demostrar la ausencia de sesgo antes del despliegue, no después de que el daño se ha producido. Williamson y Prybutok (2024) enfatizan que la privacidad de los datos es una condición previa a cualquier discusión sobre equidad: sin protección robusta, la

predicción policial se convierte en vigilancia masiva, erosionando derechos fundamentales con independencia de la precisión del algoritmo.

Zhang (2024), a partir de un caso de predicción de Alzheimer, demuestra que la equidad algorítmica es alcanzable cuando los modelos se diseñan desde el inicio con miras a la diversidad poblacional, no cuando se intenta adaptarlos a posteriori. Esta conclusión es generalizable al ámbito de la justicia: la equidad requiere modelos construidos "desde abajo" con representación de las comunidades afectadas, no modelos genéricos del Norte Global adaptados superficialmente a contextos locales. En síntesis, el análisis sectorial revela que la ética en IAG no es una abstracción filosófica: en salud, el sesgo se traduce en dolor y daño evitable; en justicia, en estigma y privación de libertad injustificados. La sostenibilidad ética de la IAG dependerá no de su potencia computacional, sino de la robustez de sus marcos de equidad, gobernanza y responsabilidad sectorial.

2.3 Marco Conceptual

A continuación, se definen los conceptos operativos fundamentales de esta investigación, derivados del marco teórico expuesto y ordenados de lo general a lo específico. Estas definiciones no son meramente descriptivas, sino que establecen el alcance y los límites analíticos con los que cada término se utiliza a lo largo del estudio.

Inteligencia Artificial Generativa (IAG)

Subcampo de la inteligencia artificial centrado en la creación de contenido nuevo y original texto, imágenes, audio, código, video a partir del aprendizaje de patrones estadísticos en datos existentes. A diferencia de la IA discriminativa, que clasifica entradas en categorías predefinidas, la IAG produce artefactos sintéticos que imitan las características estadísticas de sus datos de entrenamiento sin constituir copias directas. Los principales paradigmas arquitectónicos incluyen los Transformers, las Redes Generativas Adversariales (GANs), los Modelos de Difusión y los Autoencoders Variacionales (VAEs) (Martínez-Rolán et al., 2025).

Grandes Modelos de Lenguaje (LLMs)

Subcategoría de la IAG especializada en la comprensión y generación de texto en lenguaje natural. Los LLMs son modelos basados en la arquitectura Transformer entrenados con corpus de texto masivos que desarrollan capacidades emergentes de

razonamiento, traducción, síntesis y generación de texto. Sus implicaciones éticas son especialmente significativas dado su potencial para influir en la opinión pública, mediar en interacciones sociales y automatizar tareas que hasta ahora requerían juicio humano experto (Henao Henao y Hernández Mora, 2025).

Sesgo Algorítmico

Desviación sistemática en los resultados de un algoritmo que produce salidas injustas, inequitativas o incorrectas para ciertos grupos demográficos, culturas o contextos. En la IAG, se manifiesta en la generación de contenido que perpetúa estereotipos, subrepresenta a minorías o refleja perspectivas culturales sesgadas. Puede originarse en los datos de entrenamiento (sesgo de representación), en el diseño del algoritmo (sesgo de medición), en los procesos de evaluación (sesgo de evaluación) o en el contexto de implementación (sesgo de implementación) (Pessach y Shmueli, 2022).

Equidad Algorítmica (Fairness)

Principio normativo que busca garantizar que los resultados de un sistema de IA sean justos para todos los individuos y grupos afectados, libre de prejuicio o favoritismo basado en características sensibles inherentemente protegidas como raza, género, edad, religión u orientación sexual. La equidad algorítmica no tiene una definición única universalmente aceptada: distintas métricas paridad demográfica, igualdad de oportunidad, calibración capturan aspectos diferentes y pueden entrar en conflicto entre sí, requiriendo decisiones normativas explícitas sobre cuál priorizar en cada contexto (Li et al., 2023).

Calidad de Datos

Conjunto de propiedades que determinan la aptitud de un conjunto de datos para su propósito previsto, incluyendo las dimensiones de precisión, completitud, consistencia, actualidad y representatividad. En la IAG, la calidad de los datos de entrenamiento es un determinante crítico de la equidad, confiabilidad y generalización de los modelos resultantes. Los problemas de calidad pueden originar o amplificar sesgos algorítmicos, comprometer la seguridad de los sistemas y limitar su aplicabilidad a contextos o poblaciones no adecuadamente representadas (de Andrade et al., 2026).

IA Verde (Green AI)

Paradigma de investigación y desarrollo que prioriza la eficiencia computacional y la reducción del impacto ambiental de los modelos de IA, sin comprometer necesariamente su rendimiento. Se contrapone a la "IA Roja", que maximiza la precisión sin considerar el costo energético o las emisiones de carbono. La IA Verde abarca estrategias a nivel de algoritmo arquitecturas eficientes, pruning, cuantización, destilación, a nivel de hardware chips especializados, computación en el borde y a nivel sistémico fuentes de energía renovable para los centros de datos (Schwartz et al., 2020; Verdecchia et al., 2023).

IA Confiable (Trustworthy AI)

Sistema de IA que satisface tres condiciones fundamentales a lo largo de su ciclo de vida: es legal (cumple la normativa aplicable), ético (respeta los derechos humanos y los principios morales establecidos) y robusto (opera de forma segura, fiable y consistente bajo condiciones reales de uso). Es el marco operativo adoptado por reguladores como la Unión Europea y abarca dimensiones como la transparencia, la explicabilidad, la privacidad, la responsabilidad y la equidad. La IA Confiable no es un estado final, sino un proceso continuo que requiere auditoría, monitoreo y actualización a lo largo de la vida del sistema (Kaur et al., 2022; Li et al., 2023).

Gobernanza de la IA

Conjunto de mecanismos formales e informales regulaciones, estándares, normas, incentivos, prácticas organizacionales mediante los cuales las sociedades ejercen control sobre el desarrollo, despliegue y uso de sistemas de IA. La gobernanza efectiva opera en múltiples niveles internacional, nacional, sectorial, organizacional y requiere la participación de múltiples actores gobiernos, empresas, academia, sociedad civil. Su objetivo es asegurar que los sistemas de IA contribuyan al bienestar humano mientras se minimizan sus riesgos y consecuencias negativas (Papagiannidis et al., 2025).

Equidad en Salud Algorítmica

Principio que exige que los sistemas de IA en medicina mantengan un rendimiento equitativo en sensibilidad, especificidad y valor predictivo a través de distintos subgrupos demográficos definidos por variables sensibles como raza, género y estatus socioeconómico. Su aplicación requiere no solo métricas técnicas desagregadas, sino

también validación en contextos clínicos diversos y colaboración con expertos en equidad en salud y medicina social (Chen, R. J. et al., 2023; Abràmoff et al., 2023).

Policía Predictiva

Metodología de aplicación de la ley que utiliza algoritmos de análisis de datos para anticipar la ubicación geográfica de futuros delitos o la probabilidad de que un individuo participe en actividades delictivas, basándose en patrones históricos de datos policiales. Su principal riesgo estructural es la generación de bucles de retroalimentación donde los sesgos del sistema se confirman y amplifican a través de la propia actuación policial que el sistema orienta (Berk, 2021; Alikhademi et al., 2022).

Crisis de Sobreconfianza

Fenómeno psicológico-ético en la interacción humano-máquina por el cual los profesionales clínicos, agentes del orden, jueces depositan una confianza excesiva y acrítica en las recomendaciones del sistema de IA, asumiendo que la capacidad computacional implica infalibilidad. Este fenómeno conduce a una reducción del escrutinio humano necesario y puede amplificar los efectos de los errores algorítmicos al privar de la supervisión humana que podría detectarlos (Berkowitz et al., 2026; Selten et al., 2023).

Fallo de Transporte Algorítmico

Deterioro del rendimiento de un modelo de IA cuando se aplica en un contexto geográfico o demográfico diferente al de su entrenamiento original, debido a discrepancias en las distribuciones locales de datos. Constituye un riesgo crítico para la adopción global de IAG en contextos locales, particularmente en Latinoamérica, donde los modelos desarrollados en el Norte Global pueden fracasar debido a diferencias genéticas, epidemiológicas, lingüísticas y de registro de datos (Calahorrano Latorre, 2025; Goetz et al., 2024).

Mitigación de Sesgos

Conjunto de estrategias técnicas, metodológicas y organizacionales orientadas a identificar, medir y reducir los sesgos en sistemas de IA a lo largo de su ciclo de vida. Las estrategias pueden aplicarse en la fase de pre-procesamiento (antes del entrenamiento, actuando sobre los datos), en la fase de en-procesamiento (durante el entrenamiento, modificando el algoritmo) o en la fase de post-procesamiento (después del entrenamiento,

ajustando las salidas del modelo). La efectividad de las estrategias varía según el tipo de sesgo, el dominio de aplicación y los recursos disponibles, y ninguna estrategia única es suficiente para garantizar la equidad en todos los contextos (Skaiky et al., 2025; Rajawat et al., 2024).

Huella de Carbono Digital

Cantidad total de emisiones de gases de efecto invernadero generadas directa e indirectamente por las actividades computacionales, medida en toneladas de CO_2 equivalente. En la IAG, incluye las emisiones del entrenamiento de modelos, de la inferencia en producción y del ciclo de vida del hardware. Su cuantificación es un requisito previo para el desarrollo de estrategias de IA Verde efectivas y para la alineación del desarrollo tecnológico con los Objetivos de Desarrollo Sostenible (Budennyy et al., 2022).

2.4 Síntesis del Marco Teórico

La revisión de los antecedentes, el marco teórico y el marco conceptual revela un campo de investigación marcado por tensiones fundamentales e interrelacionadas que articulan la agenda de la IA Responsable en el presente y en el futuro próximo.

La primera tensión es entre el potencial transformador de la IAG y los riesgos que introduce. Las capacidades sin precedentes para la generación de contenido, la automatización de tareas cognitivas complejas y la aceleración del conocimiento científico conviven con riesgos igualmente significativos de perpetuación de desigualdades, erosión de la privacidad, desinformación y deterioro ambiental. Esta coexistencia de promesa y peligro define la ambivalencia constitutiva de la IAG como tecnología socio-técnica.

La segunda tensión es entre la escala necesaria para el rendimiento y la sostenibilidad necesaria para la viabilidad a largo plazo. Los modelos más grandes y capaces requieren inversiones computacionales y energéticas que crecen de manera más que lineal con el tamaño del modelo, lo que plantea interrogantes fundamentales sobre la sostenibilidad de la trayectoria actual de desarrollo de la IAG en un contexto de emergencia climática global. La IA Verde emerge no como una opción deseable sino como una condición de viabilidad.

La tercera tensión es entre la universalidad pretendida de los modelos y la particularidad de sus datos de entrenamiento. Los LLMs y otros modelos de IAG son

entrenados con datos que reflejan perspectivas culturales, lingüísticas y demográficas específicas predominantemente anglosajonas, occidentales y de clase media con acceso a internet, pero se despliegan en contextos globales enormemente diversos. Esta brecha entre contextos de entrenamiento y contextos de uso es una fuente estructural de sesgos, fallos de generalización y fallo de transporte algorítmico que afecta de manera desproporcionada al Sur Global.

La cuarta tensión, finalmente, es entre la velocidad del desarrollo tecnológico y la capacidad de los marcos regulatorios, éticos y sociales para adaptarse. La investigación sobre ética algorítmica, calidad de datos y IA Verde avanza rápidamente, pero la industria continúa desplegando sistemas de IAG a una velocidad que dificulta la implementación sistemática de las mejores prácticas identificadas por la academia y los reguladores. El AI Act europeo representa un primer paso significativo, pero su implementación efectiva seguirá siendo un desafío de primer orden en los próximos años.

La intersección de los ocho pilares teóricos desarrollados en este marco justicia algorítmica y gobernanza, sesgos en modelos generativos, sesgos específicos en LLMs, calidad de los datos, sostenibilidad ambiental, implicaciones en salud, justicia algorítmica en seguridad pública, y estrategias de mitigación define la frontera actual de la investigación en IA Responsable y constituye el andamiaje analítico desde el cual este estudio examina la literatura científica disponible. La comprensión de estas tensiones no es solo un ejercicio académico: es una condición necesaria para el desarrollo de sistemas de IAG que sean, simultáneamente, capaces, equitativos, confiables y sostenibles. Los sectores de salud y justicia, donde los errores algorítmicos se traducen en daño físico, privación de libertad y perpetuación de desigualdades, constituyen el laboratorio más exigente y el más urgente para esta agenda de investigación.

3 DISEÑO METODOLÓGICO

El presente capítulo expone el diseño metodológico que orientó la investigación sobre la Inteligencia Artificial Generativa (IAG): sus implicaciones éticas, patrones de sesgo algorítmico y retos de sostenibilidad. La elección del enfoque, el tipo de estudio, la estrategia de búsqueda y los procedimientos de análisis responden a la naturaleza del problema y a la pregunta de investigación planteada: ¿cómo influyen los sesgos cognitivos, la calidad de

los datos de entrenamiento y el impacto ambiental en el desarrollo ético y sostenible de la IAG? Dado que el objetivo central es examinar, sintetizar y valorar el estado del conocimiento científico disponible y no producir datos primarios ni experimentar con sistemas de IA, el diseño seleccionado es el de revisión sistemática de literatura, ejecutada bajo el protocolo PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses).

3.1 Tipo y Enfoque de la Investigación

La investigación es de tipo documental, con un componente analítico-comparativo que permite identificar patrones, tensiones y convergencias a través de un corpus bibliográfico diverso y representativo. El enfoque es mixto en sentido metodológico: cuantitativo en la sistematización del proceso de búsqueda y selección conteo de registros, distribución temática y temporal, y cualitativo en el análisis de contenido, la interpretación de los hallazgos y la síntesis narrativa de la evidencia disponible.

Se descartaron el diseño experimental y el casi-experimental porque la investigación no manipula variables ni controla grupos. Igualmente, se descartó la encuesta o el estudio de caso primario porque el objeto de análisis no son actores organizacionales específicos, sino la producción científica sobre un fenómeno tecnológico emergente. La revisión sistemática de literatura es el diseño canónico para investigaciones cuyo propósito es valorar el estado del arte en un campo en rápida evolución y establecer conclusiones basadas en evidencia acumulada, condiciones que se cumplen plenamente en el estudio de la IAG ética y sostenible.

El alcance de la investigación es descriptivo-analítico: descriptivo porque caracteriza el estado actual del conocimiento sobre los tres ejes (sesgos, ética/gobernanza y sostenibilidad); analítico porque descompone los fenómenos en sus dimensiones constitutivas, examina las relaciones entre ellos y evalúa críticamente las estrategias de mitigación y gobernanza propuestas en la literatura. No pretende ser predictivo ni prescriptivo, aunque sus conclusiones y recomendaciones ofrecen orientaciones para la práctica y la política pública.

3.2 Protocolo de Revisión Sistemática: PRISMA 2020

El protocolo PRISMA 2020 fue adoptado como estándar metodológico para garantizar la transparencia, reproducibilidad y rigor del proceso de revisión. PRISMA 2020 es el estándar más ampliamente aceptado en la literatura científica para la conducción y reporte de revisiones sistemáticas, tanto en ciencias de la salud como en ciencias de la computación y ciencias sociales. Su estructura en cuatro fases identificación, cribado, elegibilidad e inclusión permite documentar con precisión cada decisión de inclusión o exclusión y asegura que los resultados sean replicables por investigadores independientes.

La adopción de este protocolo responde, además, a la escala y diversidad temática del corpus analizado. Con cinco grupos temáticos de búsqueda que atraviesan disciplinas

tan diversas como la informática, la bioética, el derecho, la ciencia política y la ingeniería ambiental, era esencial contar con un procedimiento estandarizado que garantizara la coherencia de los criterios de selección a lo largo de todo el proceso.

3.3 Estrategia de Búsqueda Bibliográfica

3.3.1 Período y Bases de Datos Consultadas

La búsqueda bibliográfica se ejecutó entre septiembre y diciembre de 2025 en cuatro bases de datos académicas de alto impacto internacional, seleccionadas por su cobertura complementaria de las disciplinas involucradas en la investigación:

- IEEE Xplore Digital Library: especializada en ingeniería eléctrica, electrónica y ciencias de la computación, con amplia cobertura de actas de conferencias de primer nivel como IEEE, ACM y NeurIPS.
- Research EBSCO: base de datos multidisciplinaria con acceso a revistas en ciencias sociales, humanidades, salud y gestión, pertinente para los ejes de ética, gobernanza y aplicaciones sectoriales.
- Scopus (Elsevier): la base de datos de citas científicas más amplia del mundo, con indexación de más de 25.000 revistas activas en todas las disciplinas, que permite verificar el cuartil de impacto (Q1/Q2) de las publicaciones.
- Web of Science (Clarivate): referencia estándar para el análisis bibliométrico y la evaluación de la calidad de las publicaciones, con especial fortaleza en ciencias naturales y ciencias de la computación.

La complementariedad de estas bases permite una cobertura amplia y equilibrada del corpus científico relevante, reduciendo el riesgo de sesgo de publicación derivado de la dependencia de una única plataforma.

3.3.2 Grupos Temáticos y Ecuaciones de Búsqueda

Las referencias se organizaron en cinco grupos temáticos que corresponden a los ejes centrales de la investigación. Para cada grupo se diseñaron ecuaciones de búsqueda específicas utilizando operadores booleanos (OR, AND) y vocabulario controlado, combinando términos en inglés y español para maximizar la recuperación de literatura relevante:

Tabla 1. Grupos temáticos de búsqueda y ecuaciones principales utilizadas en la revisión sistemática.

Grupo Temático	Eje de la Investigación	Ecuaciones de Búsqueda Principales
AI Sostenible	Sostenibilidad ambiental de la IA	"Green AI" OR "carbon footprint machine learning" OR "sustainable AI" OR "energy consumption deep learning"
Calidad de Datos	Datos de entrenamiento	"Dataset bias" OR "data quality AI" OR "bias in training data" OR "electronic health records AI quality"
Ética y Gobernanza	Marcos éticos y regulación	"Responsible AI" OR "trustworthy AI" OR "AI ethics frameworks" OR "EU AI Act" OR "algorithmic governance"
Salud y Justicia	Sesgos sectoriales en salud y justicia	"AI healthcare bias" OR "predictive policing" OR "algorithmic fairness health" OR "AI criminal justice"
Sesgo LLM	Sesgos en modelos de lenguaje	"LLM bias" OR "large language model fairness" OR "gender bias NLP" OR "political bias language models"

Nota. La tabla resume la estrategia de recuperación bibliográfica para los 104,660 registros iniciales. Los términos combinan vocabulario controlado en inglés y español para asegurar la cobertura técnica. Elaboración propia (2026).

La selección de cinco grupos temáticos responde a la estructura conceptual de la investigación: los grupos de Ética y Gobernanza y Sesgo LLM cubren el eje de las implicaciones éticas y los sesgos algorítmicos en los LLMs; el grupo de Calidad de Datos aborda la dimensión de los datos de entrenamiento como origen estructural de los sesgos; el grupo de Salud y Justicia aporta la perspectiva de los impactos sectoriales en contextos críticos; y el grupo de AI Sostenible cubre el eje de la sostenibilidad ambiental y la huella de carbono.

3.4 Criterios de Inclusión y Exclusión

3.4.1 Criterios de Inclusión

Para ser incluida en el corpus final, cada referencia debía satisfacer simultáneamente todos los criterios siguientes:

- Rango temporal: publicaciones entre 2016 y 2026 con texto final publicado y revisado por pares. Se admitieron excepcionalmente referencias anteriores a 2016 cuando se tratara de obras fundacionales explícitamente justificadas en el marco teórico como la teoría rawlsiana de la justicia o los orígenes del paradigma de sostenibilidad cuya relevancia teórica trasciende el período contemporáneo.
- Indexación y calidad: artículos publicados en revistas clasificadas en los cuartiles Q1 o Q2 según Scimago Journal Ranking (SJR) o Journal Citation Reports (JCR); actas de conferencias de primer nivel con revisión por pares (ACM FAccT, IEEE, NeurIPS, ICML); libros de editoriales académicas reconocidas (MIT Press, Oxford University Press, Yale University Press, Springer); o documentos oficiales de organismos internacionales de referencia (Unión Europea, Naciones Unidas, UNESCO, IEEE Standards Association).
- Pertinencia temática: estudios empíricos, revisiones sistemáticas o marcos conceptuales directamente relacionados con los tres ejes centrales de la investigación. Se aceptaron estudios de aplicación sectorial en salud, agricultura, justicia, ciudades inteligentes cuando ilustran con evidencia empírica la escala de adopción de la IA o sus impactos reales en contextos específicos.
- Idioma: inglés o español, los dos idiomas en los que se concentra la producción científica relevante para los temas de investigación.
- Accesibilidad: texto completo accesible a través de las plataformas de bases de datos institucionales (CRAI USTA Digital) o con DOI verificable en el momento de la búsqueda.

3.4.2 Criterios de Exclusión

Fueron excluidas las referencias que presentaban alguna de las siguientes características:

- Revistas predatorias o publicaciones sin proceso de revisión por pares documentado, verificadas contra la Beall's List 2024 y confirmadas por ausencia de ISSN verificable o indexación en Scopus o WoS.
- Publicaciones sin identificación de autores o sin afiliación institucional verificable, lo que impide evaluar la trayectoria y credibilidad académica de los autores.
- Hardware puro o infraestructura de red sin vínculo conceptual con los ejes de la monografía: estudios sobre aceleradores para álgebra matricial, fotónica computacional, memorias RRAM y similares que no vinculan sus desarrollos con implicaciones éticas, de sesgo o ambientales.
- Actas de talleres (workshops) o conferencias menores de alcance regional sin indexación en Scopus o Web of Science, que no ofrecen las garantías de rigor del proceso de revisión por pares propio de las conferencias de primer nivel.

- Estudios con metodología no claramente definida o con resultados no generalizables más allá del caso específico analizado.
- Duplicados: cuando el mismo trabajo estaba disponible en varias versiones preprint más publicación final, se retuvo únicamente la versión publicada y revisada por pares.
- Estudios que utilizan la IA como herramienta instrumental para resolver otros problemas sin abordar sus implicaciones éticas, de sesgo o ambientales como modelos de IA para la optimización de edificios o la gestión de ganadería sin análisis de impacto.

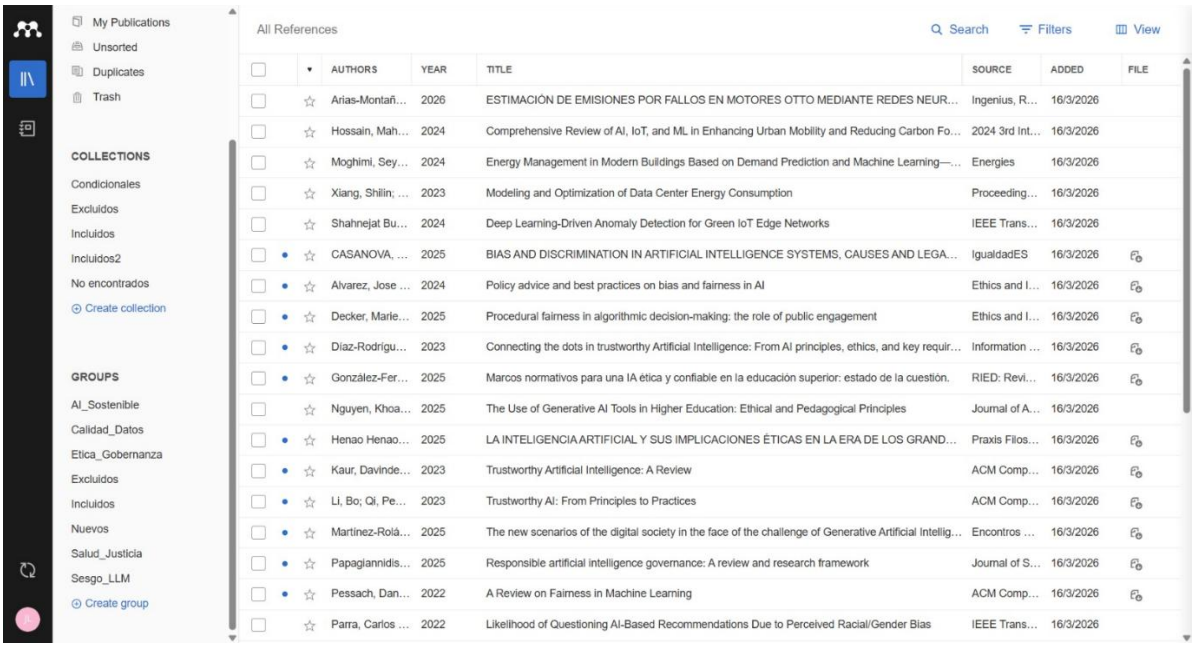
3.5 Proceso de Selección: Flujo PRISMA 2020

El proceso de selección se desarrolló en cuatro fases sucesivas, conforme al diagrama de flujo PRISMA 2020, garantizando que cada decisión de inclusión o exclusión fuera documentada con sus motivos específicos.

3.5.1 Fase 1: Identificación

Se identificaron inicialmente 104,660 registros en las cuatro bases de datos consultadas (EBSCO: 7,432; Web of Science: 12,927; Scopus (Elsevier): 10,860; IEEE Xplore: 73,441). Tras la eliminación de duplicados e irrelevantes, se obtuvieron 293 registros únicos distribuidos en los cinco grupos temáticos de búsqueda, gestionados mediante el gestor bibliográfico Mendeley. La distribución de registros seleccionados para revisión por grupo fue la siguiente: AI Sostenible (55 registros), Calidad de Datos (46 registros), Ética y Gobernanza (57 registros), Salud y Justicia (79 registros) y Sesgo LLM (56 registros). La mayor concentración de registros en el grupo de Salud y Justicia refleja la abundancia de literatura empírica disponible sobre sesgos algorítmicos en estos sectores, particularmente en el ámbito de la IA sanitaria.

Figura 1. Interfaz del gestor de referencias Mendeley con la biblioteca del estudio.

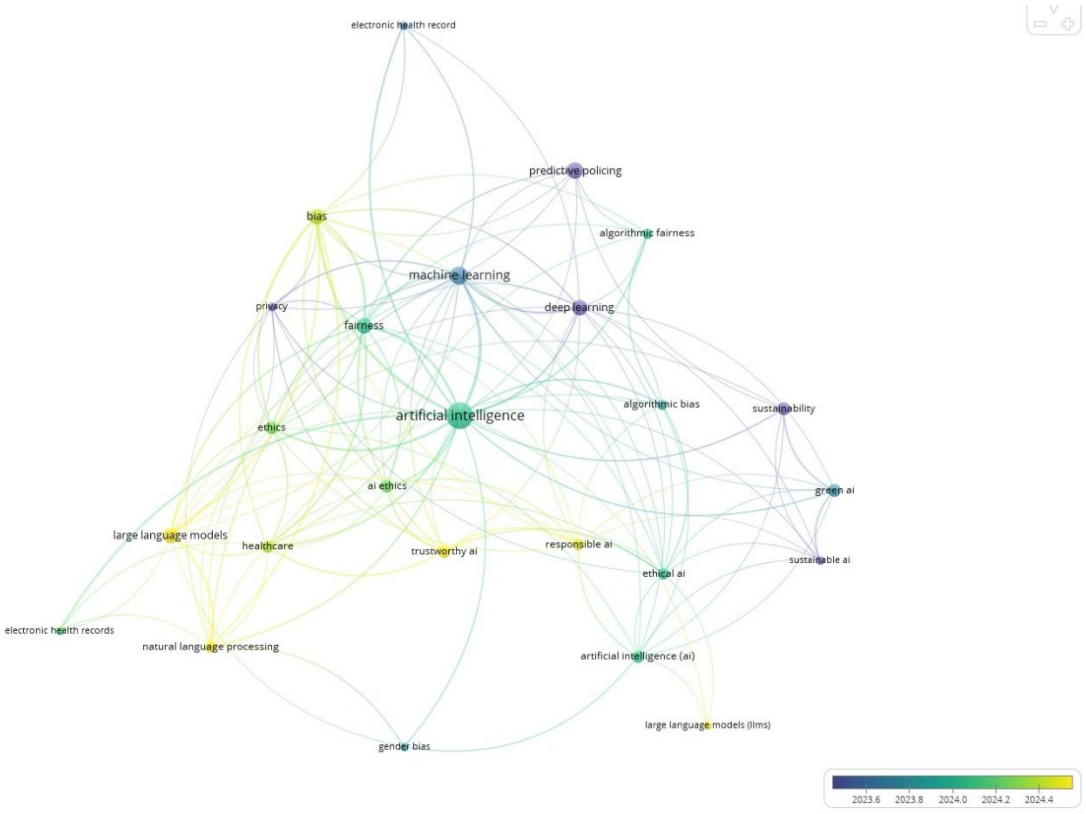


Nota. La imagen muestra la organización de las fuentes bibliográficas mediante el software Mendeley. Elaborada por Vergel y López, 2026.

3.5.2 Fase 2: Cribado (Screening)

En la fase de cribado se revisaron los títulos, resúmenes e información de indexación de los 293 registros únicos identificados. Los 293 registros fueron remitidos en su totalidad a revisión de texto completo, dado que el filtro por título y resumen confirmó su pertinencia temática general con los ejes de la investigación. No se descartaron registros en esta fase preliminar.

Figura 2. Visualización de red (Overlay Visualization) de palabras clave en VOSviewer.



Nota. Mapa bibliométrico que muestra la densidad y relación entre conceptos como "artificial intelligence", "ethics" y "machine learning". Elaboración propia a partir de datos exportados de Mendeley (2026)

3.5.3 Fase 3: Elegibilidad

En la fase de elegibilidad se procedió a la revisión del texto completo de los 293 registros candidatos. Este análisis más profundo permitió evaluar cada registro frente a la totalidad de los criterios de inclusión y exclusión. Se excluyeron 163 registros por las siguientes razones: fuentes predatorias o sin revisión por pares documentada, hardware o infraestructura de red sin vínculo con los ejes de la investigación, IA utilizada como

herramienta sin análisis ético o ambiental, actas de talleres o conferencias menores sin indexación en Scopus o WoS, metodología no claramente definida o resultados no generalizables, duplicados (versiones preprint frente a publicación final), y ausencia de indexación Q1/Q2 verificable. Los 130 registros restantes cumplieron plenamente todos los criterios de inclusión y conforman el corpus final de la revisión.

3.5.4 Fase 4: Inclusión

Las 130 fuentes incluidas fueron organizadas en los cinco grupos temáticos originales y mapeadas a las secciones específicas del borrador de la monografía donde aportan evidencia. Este mapeo bidireccional de la sección a la referencia y de la referencia a la sección permite asegurar que cada afirmación relevante del texto está respaldada por evidencia del corpus y que ninguna referencia incluida queda sin uso efectivo en el documento.

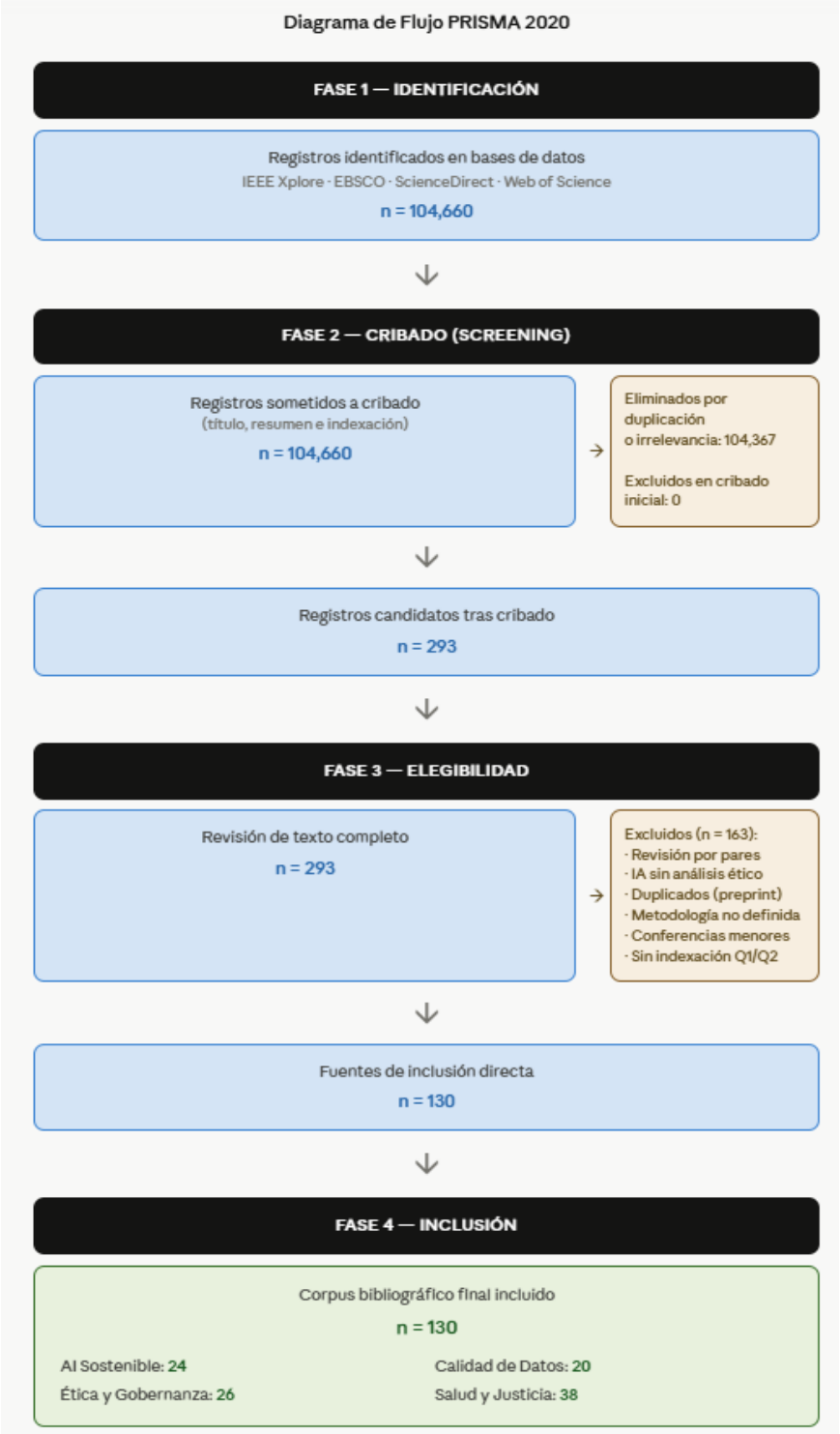
La distribución final del corpus de 130 fuentes por grupo temático quedó de la siguiente manera:

Tabla 2. Resultados de la selección bibliográfica por grupo temático siguiendo el protocolo PRISMA 2020.

Grupo Temático	Total Revisados	✓ Incluidos	✗ Excluidos
AI Sostenible	55	24	31
Calidad de Datos	46	20	26
Ética y Gobernanza	57	26	31
Salud y Justicia	79	38	41
Sesgo LLM	56	22	34
TOTAL	**293**	**130**	**163**

Nota. El desglose detalla la transición de la fase de elegibilidad a la inclusión final. Elaboración propia (2026).

Figura 3. Diagrama de flujo PRISMA 2020 del proceso de búsqueda y selección bibliográfica.



Nota. Los datos reflejan el proceso de depuración desde 104,660 registros iniciales hasta la selección de 130 artículos finales. Elaboración propia (2026).

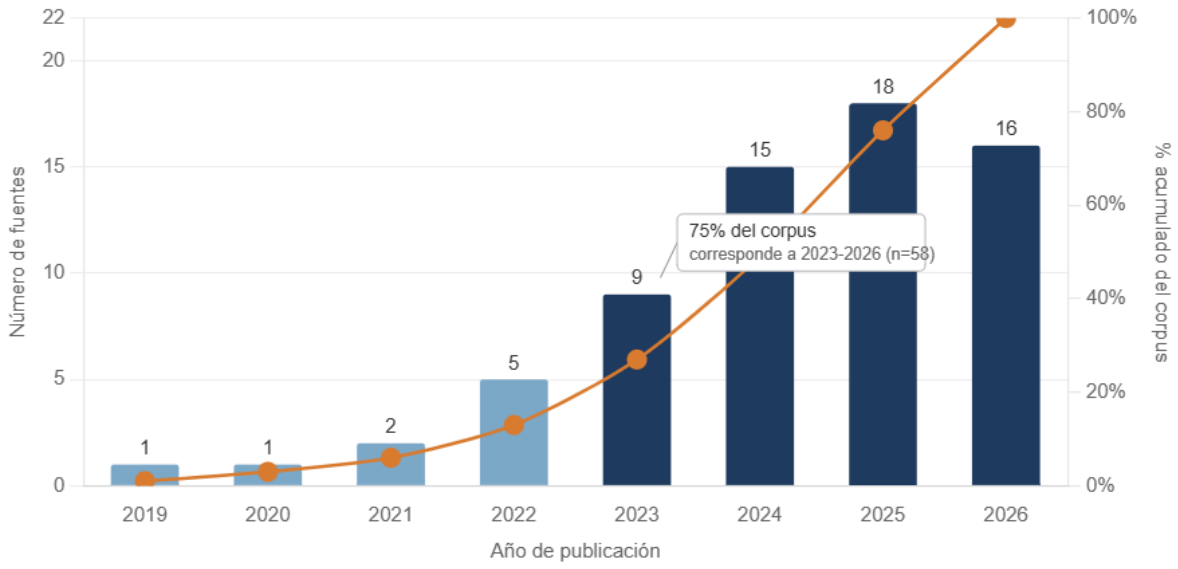
3.6 Caracterización del Corpus Final

El corpus final de 130 fuentes presenta características que garantizan su rigor, actualidad y representatividad temática para los objetivos de la investigación.

3.6.1 Distribución Temporal

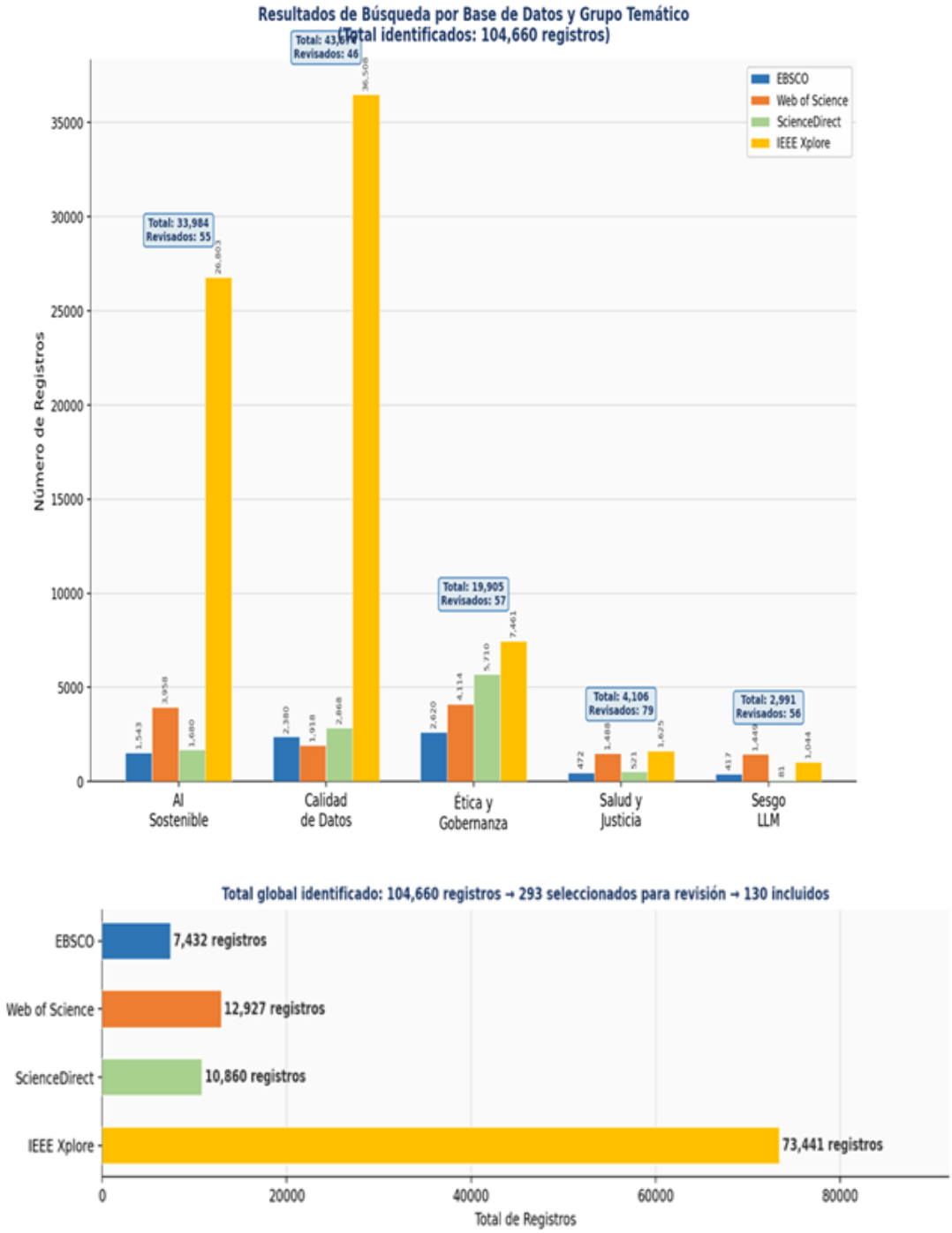
El 75% de las fuentes corresponde al período 2023-2026, lo que garantiza la máxima actualidad del corpus en un campo de rápida evolución como la IAG. Esta concentración en publicaciones recientes es deliberada y metodológicamente justificada: dado que los grandes modelos de lenguaje como objeto de estudio de sesgos y sostenibilidad son un fenómeno esencialmente posterior a 2019, incluir literatura anterior a ese año solo tendría sentido para obras fundacionales de carácter teórico. El período 2024-2026 aporta el mayor número de fuentes (aproximadamente el 52% del total), reflejando la explosión de investigación sobre ética y sesgos en IAG que ha acompañado la masificación de herramientas como ChatGPT, Gemini y Claude en el contexto académico y profesional global.

Figura 4. Distribución temporal fuentes bibliográficas.



Nota. Análisis de densidad temporal derivado de los metadatos obtenidos en la fase de identificación. Visualización generada con IA mediante el análisis de frecuencias por año del corpus final (n=130). Elaboración propia (2026).

Figura 5. Distribución de registros identificados por base de datos y grupo temático (N = 104,660 → 293 seleccionados).



Nota. Análisis comparativo de la representatividad temática y procedencia de registros. El grafo fue generado con IA a partir del procesamiento de los datos estadísticos obtenidos en el Diagrama PRISMA (Figura 3). Elaboración propia (2026).

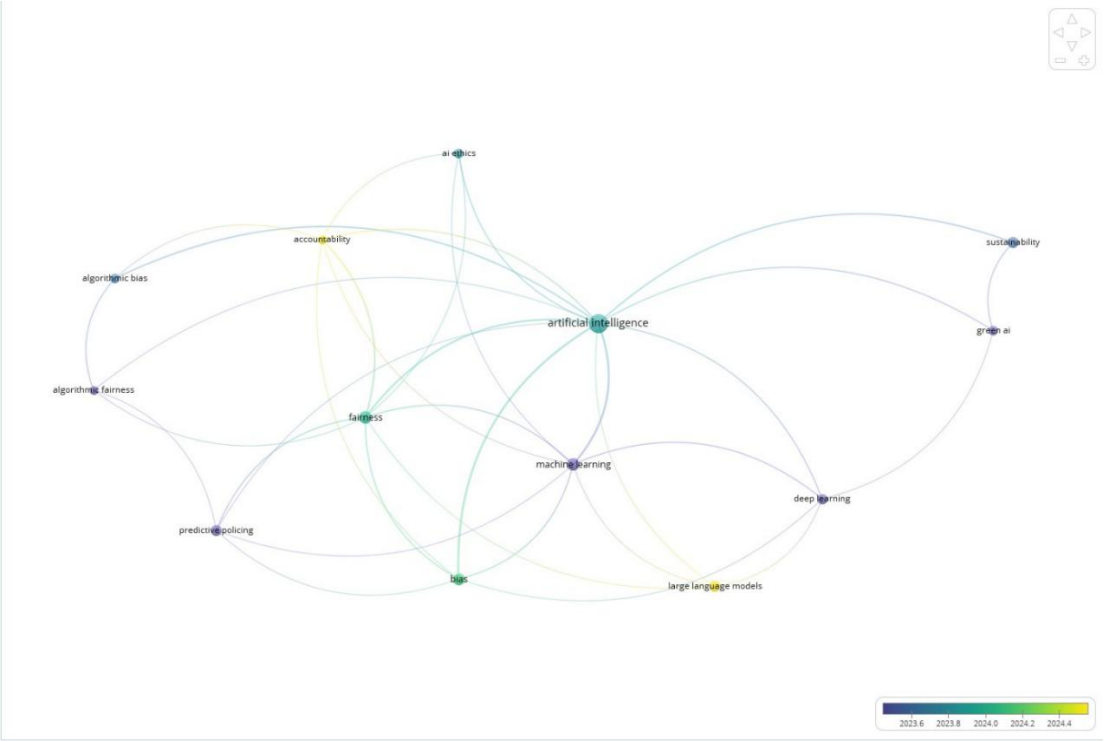
3.6.2 Análisis Bibliométrico: Mapa de Co-ocurrencia (VOSviewer)

Para complementar nuestra revisión con una visión más amplia del campo, usamos **VOSviewer** para mapear cómo se conectan los conceptos clave en los 293 registros iniciales. A diferencia del análisis detallado, este mapa nos permite ver 'la foto completa' de la investigación en IA generativa.

Diseñamos la búsqueda usando términos como 'bias', 'ethics' y 'sustainability' en títulos y resúmenes para no dejar fuera ningún estudio relevante. Al construir el mapa, filtramos los términos que aparecían menos de 5 veces; así limpiamos el 'ruido' y nos quedamos solo con los conceptos que realmente mueven la aguja en el sector.

El resultado son varios grupos de temas (clústeres) que muestran claramente las líneas de investigación actuales: desde los sesgos y la equidad, hasta la ética y el impacto ambiental. Lo más interesante es ver qué tan cerca están unos de otros, lo que demuestra que la IA generativa no se estudia de forma aislada, sino como un cruce entre lo técnico, lo social y lo ambiental.

Figura 6. Mapa de co-ocurrencia de términos generado con VOSviewer a partir del conjunto ampliado de registros (n = 130).



Nota. Visualización de red (Overlay Visualization) que muestra la relación y evolución temporal de palabras clave. Se observa el desplazamiento desde conceptos técnicos hacia la ética y gobernanza en los años más recientes (2024.4). Elaboración propia (2026).

3.6.3 Distribución por Tipo de Fuente

El 79% de las fuentes son artículos publicados en revistas Q1/Q2 o actas de conferencias de primer nivel (ACM, IEEE, Nature, JAMA Network Open, npj Digital Medicine), lo que garantiza el rigor del proceso de revisión por pares del corpus. Las revistas de mayor frecuencia de aparición incluyen ACM Computing Surveys, Nature Biomedical Engineering, Ethics and Information Technology, IEEE Transactions on Affective Computing y el Journal of the American Medical Informatics Association (JAMIA). Esta distribución refleja la naturaleza interdisciplinaria de la investigación, que requiere fuentes tanto de informática y aprendizaje automático como de bioética, derecho, medicina y ciencias ambientales.

Los libros y reportes institucionales (7%) corresponden a obras de referencia canónicas en el campo, mientras que una pequeña proporción de fuentes (4%) corresponde a documentos oficiales de organismos internacionales como las directrices de la Unión Europea para IA Confiable cuya autoridad normativa justifica su inclusión independientemente de si son artículos revisados por pares.

3.6.4 Distribución Geográfica y Lingüística

Aunque la mayoría de las fuentes están escritas en inglés lo que refleja la dominancia del inglés como lengua franca de la publicación científica, se incorporaron deliberadamente fuentes en español que aportan perspectivas latinoamericanas y europeas de habla hispana sobre la ética y la regulación de la IA (González-Fernández et al., 2025; Henao Henao y Hernández Mora, 2025; Calahorrano Latorre, 2025; Sarrionandia et al., 2025; Melo, 2024). Esta inclusión es metodológicamente significativa porque permite identificar convergencias y divergencias entre el debate global y las preocupaciones específicas del contexto latinoamericano, donde los marcos regulatorios y las capacidades institucionales para la gobernanza de la IA presentan características propias.

3.7 Procedimiento de Análisis de la Literatura

El análisis de las 130 fuentes incluidas se desarrolló en tres fases sucesivas y complementarias, siguiendo los principios del análisis cualitativo de contenido y la síntesis narrativa sistemática.

3.7.1 Fase 1: Categorización Temática

En la primera fase, cada fuente fue leída en su totalidad y clasificada según su aporte principal a uno o más de los tres ejes de investigación: sesgos algorítmicos, ética y gobernanza, y sostenibilidad ambiental. Se elaboró una matriz de análisis en la que se registró, para cada fuente: el eje o ejes a los que contribuye, el tipo de contribución (empírica, teórica o de revisión), los principales conceptos y hallazgos relevantes, las estrategias de mitigación o marcos propuestos, y las limitaciones señaladas por los propios

autores. Esta matriz permitió identificar las referencias más relevantes para cada sección del texto y anticipar las tensiones y convergencias que emergerían en el análisis posterior.

3.7.2 Fase 2: Análisis Comparativo

En la segunda fase se realizó una comparación sistemática entre las fuentes de un mismo eje temático para identificar: acuerdos y consensos en la literatura (por ejemplo, el reconocimiento generalizado de que los sesgos en IAG son sistémicos y no meramente técnicos); debates y controversias activas (como la discusión sobre qué métrica de equidad debe priorizarse cuando diferentes definiciones entran en conflicto); brechas en el conocimiento disponible (por ejemplo, la escasez de estudios sobre sesgos en LLMs en idiomas distintos al inglés o en contextos latinoamericanos); y evoluciones conceptuales (como la transición del concepto de 'ética en IA' hacia marcos más operativos de 'IA Confiable' y, más recientemente, hacia marcos regulatorios vinculantes).

El análisis comparativo también incluyó la identificación de estudios con metodologías particularmente robustas revisiones sistemáticas con metaanálisis, estudios empíricos con grandes volúmenes de datos, ensayos clínicos aleatorizados sobre herramientas de IA médica para otorgarles mayor peso argumentativo en la síntesis final.

3.7.3 Fase 3: Síntesis Integradora

En la tercera fase se elaboró la síntesis narrativa que integra los hallazgos de las tres áreas temáticas en una argumentación coherente sobre la IAG como fenómeno socio-técnico. La síntesis narrativa es el método preferido para revisiones sistemáticas cuando la heterogeneidad de los estudios primarios en términos de diseño, población, contexto y medidas de resultado hace inviable o inapropiado un metaanálisis estadístico. En esta investigación, la heterogeneidad es intrínseca: los estudios sobre sesgos en LLMs, los estudios sobre calidad de datos clínicos y los estudios sobre huella de carbono de la IA utilizan metodologías, métricas y contextos tan distintos que solo una síntesis narrativa puede integrar sus contribuciones de manera significativa.

La síntesis se orientó explícitamente hacia la identificación de las tensiones fundamentales que atraviesan el campo entre capacidades técnicas y equidad algorítmica; entre escala del modelo y sostenibilidad; entre universalidad pretendida y particularidad de los datos y hacia la articulación de las condiciones que harían posible una IAG ética, equitativa y sostenible.

3.8 Rigor y Validez de la Revisión

La revisión sistemática adoptó varias estrategias para garantizar su rigor y controlar posibles fuentes de sesgo metodológico. En primer lugar, la triangulación de fuentes: al consultar cuatro bases de datos con coberturas complementarias, se redujo el riesgo de sesgo de indexación la posibilidad de que estudios relevantes estuvieran disponibles solo en una base de datos y no en otras. En segundo lugar, la documentación explícita de todos

los criterios de inclusión y exclusión antes del inicio de la búsqueda permite evaluar la consistencia de las decisiones tomadas a lo largo del proceso.

En tercer lugar, la aplicación rigurosa de los criterios de inclusión y exclusión durante la fase de elegibilidad, que permitió reducir el corpus de 293 registros únicos a 130 fuentes de alta calidad, garantiza la coherencia interna y la pertinencia temática del corpus final. En cuarto lugar, el mapeo explícito de cada referencia a las secciones específicas del texto en las que aporta evidencia documentado en el Anexo de Búsqueda Bibliográfica permite verificar que cada afirmación relevante del documento está respaldada por al menos una fuente del corpus y que ninguna referencia queda sin función argumentativa.

Como medida de control de sesgo de publicación la tendencia a que los estudios con resultados positivos sean más fácilmente publicados que los estudios con resultados negativos o nulos, se buscó activamente incluir fuentes que documenten limitaciones, fracasos o efectos no deseados de las estrategias de mitigación de sesgos y de los enfoques de IA Verde, además de los estudios que reportan éxitos. Esto es especialmente relevante en un campo donde el entusiasmo por las nuevas tecnologías puede generar presiones para reportar resultados favorables.

3.9 Limitaciones Metodológicas

El presente diseño metodológico presenta limitaciones que deben ser reconocidas explícitamente para contextualizar adecuadamente las conclusiones de la investigación.

La primera limitación es la ausencia de experimentación empírica propia. La investigación se basa exclusivamente en la literatura científica disponible y no genera datos primarios mediante la evaluación directa de modelos de IAG, la realización de entrevistas con desarrolladores o la medición empírica de indicadores de equidad o sostenibilidad. Esta limitación es inherente al diseño de revisión sistemática y es metodológicamente aceptable dado que el objetivo de la investigación es examinar el estado del conocimiento existente, no generar conocimiento primario nuevo. Sin embargo, implica que las conclusiones sobre la efectividad de las estrategias de mitigación están mediadas por los diseños metodológicos y los contextos de los estudios primarios revisados.

La segunda limitación es la posibilidad de sesgo lingüístico. Aunque se incluyeron fuentes en inglés y español, la gran mayoría de la literatura científica sobre IAG está en inglés, lo que significa que perspectivas de comunidades académicas que publican predominantemente en otros idiomas árabe, chino, portugués, francés podrían estar subrepresentadas. Esto es especialmente relevante para el análisis de los sesgos culturales y lingüísticos de los LLMs, donde las perspectivas de comunidades no anglófonas son particularmente valiosas.

La tercera limitación es la velocidad de evolución del campo. Dado que la IAG es un área en transformación acelerada, algunos hallazgos reportados en fuentes publicadas en 2022 o 2023 pueden haber sido superados o matizados por desarrollos posteriores. Para mitigar esta limitación, se priorizaron las fuentes más recientes (2024-2026) y se verificó, en la medida de lo posible, que los hallazgos de fuentes más antiguas no contradicen el estado del arte más reciente.

La cuarta limitación es la imposibilidad de incluir literatura gris de alta relevancia informes técnicos de empresas como OpenAI, Google DeepMind o Anthropic sobre sus propios modelos porque la ausencia de revisión por pares impide asegurar su rigor metodológico, aun cuando en algunos casos estas fuentes contienen información empírica de alto valor sobre el comportamiento real de los modelos en producción. Se optó por referenciar esta información cuando fue reportada y analizada por fuentes académicas revisadas por pares, en lugar de citarla directamente.

3.10 Consideraciones Éticas del Proceso Investigativo

La integridad académica constituyó un principio rector a lo largo de todo el proceso investigativo. Todas las fuentes fueron citadas correctamente siguiendo las normas APA 7.ª edición, sin reproducir fragmentos extensos de las publicaciones originales más allá de lo permitido por los principios de uso justo y citación académica. Los hallazgos de los estudios primarios fueron presentados mediante paráfrasis y síntesis que respetan el sentido original de los autores sin apropiarse de su expresión literal.

La investigación no involucró experimentación con seres humanos, recolección de datos personales, intervención en sistemas de información ni acceso a datos clínicos, judiciales u otros datos sensibles. Por lo tanto, no requirió evaluación por parte de un comité de ética institucional. Sin embargo, los autores reconocen una responsabilidad ética en la representación fiel y equilibrada de la literatura revisada, incluyendo fuentes que presentan perspectivas críticas o contradictorias respecto a las posiciones mayoritarias del campo.

Finalmente, se reconoce que el propio proceso de selección bibliográfica implica decisiones que no son completamente neutrales: la elección de las bases de datos, los términos de búsqueda y los criterios de inclusión moldea el corpus que se analiza y, por ende, las conclusiones que se derivan. La documentación explícita de todas estas decisiones en este capítulo y en el Anexo de Búsqueda Bibliográfica permite a cualquier lector o investigador evaluar críticamente el proceso y replicarlo o adaptarlo para investigaciones futuras, cumpliendo así con el principio de transparencia que caracteriza a la investigación científica responsable.

4 RESULTADOS

El presente capítulo sintetiza los hallazgos derivados del análisis de las 130 fuentes del corpus bibliográfico final, organizado en cuatro grandes ejes temáticos que responden directamente a los objetivos específicos de la investigación: la tipología y patrones de sesgos identificados en la IAG, los casos documentados por sector, las estrategias de mitigación evaluadas en la literatura y el impacto ambiental cuantificado del desarrollo y operación de modelos de IAG. La presentación de los resultados sigue el orden analítico establecido en el diseño metodológico, aunque las interrelaciones entre ejes se señalan explícitamente cuando la evidencia así lo exige.

La distribución del corpus por subsección de resultados refleja la estructura temática de la investigación: 38 fuentes aportan evidencia a la sección 4.3 (mitigación), lo que indica que la producción científica reciente está orientada predominantemente hacia el desarrollo de soluciones; 29 fuentes contribuyen a la sección 4.2 (casos sectoriales); 26 fuentes informan la sección 4.1 (tipología de sesgos); y 22 fuentes sustentan la sección 4.4 (impacto ambiental). Varias fuentes aportan de manera simultánea a más de una subsección, lo que refleja la naturaleza multidimensional de los estudios sobre IA responsable.

Nota: La relación del cumplimiento de los objetivos específicos se detalla en el Anexo 1.

4.1 Tipología y Patrones de Sesgos en la Inteligencia Artificial Generativa

El análisis sistemático de la literatura revela que el sesgo en los sistemas de IAG no constituye un fenómeno singular ni técnicamente discreto, sino una familia de distorsiones interrelacionadas que emergen en distintas etapas del ciclo de vida de los modelos y que se manifiestan con intensidad y consecuencias variables según el dominio de aplicación. La revisión del corpus permitió identificar y caracterizar seis categorías principales de sesgos, cuya taxonomía integra y supera las clasificaciones precedentes propuestas en la literatura.

4.1.1 Sesgos de Representación

Los sesgos de representación constituyen la categoría más ampliamente documentada en el corpus. Ocurren cuando los conjuntos de datos de entrenamiento no reflejan adecuadamente la diversidad demográfica, cultural, lingüística o contextual de las poblaciones sobre las que el modelo operará. Su origen es estructural: refleja qué tipos de

información se producen, se registran y se digitalizan de manera sistemática en la sociedad, y qué grupos tienen capacidad y acceso para generar esa información.

Henao Henao y Hernández Mora (2025) documentan un caso paradigmático: el conjunto de datos ImageNet, durante mucho tiempo referencia estándar para el entrenamiento de modelos de visión computacional, contiene aproximadamente un 45% de imágenes provenientes de los Estados Unidos, lo que genera modelos con capacidades muy superiores para interpretar escenas y objetos del contexto norteamericano respecto de cualquier otra región del mundo. Li y Deng (2022) demuestran, mediante un experimento riguroso, que los modelos de reconocimiento de expresiones faciales entrenados con datasets de laboratorio pueden identificar a qué base de datos pertenece una imagen con una precisión del 79%, cuando el azar predicho sería del 14%; esta capacidad para discriminar el origen del dataset es evidencia directa de que los modelos aprenden las características idiosincráticas y sesgadas de cada corpus de entrenamiento en lugar de aprender representaciones verdaderamente generalizables.

En el ámbito del lenguaje, Ducel et al. (2023) realizaron un metaanálisis de la propia investigación sobre sesgos en LLMs y encontraron que el 97% de los estudios se centran exclusivamente en el idioma inglés, que el 70% de los autores trabajan en instituciones de los Estados Unidos y que el 39% de los artículos revisados involucran financiación de grandes corporaciones tecnológicas. Este sesgo de publicación implica que el conocimiento disponible sobre sesgos de representación en LLMs está en sí mismo sesgado hacia un contexto cultural y lingüístico particular, generando puntos ciegos sistemáticos sobre la experiencia de comunidades no anglófonas. Sarrionandia et al. (2025), en una investigación específicamente dirigida al sesgo territorial en modelos generativos de lenguaje, demostraron que cinco modelos prominentes incluyendo versiones de ChatGPT, Gemini y Claude presentan sesgos sistemáticos que favorecen a las regiones metropolitanas y a los países con mayor presencia digital frente a territorios más pequeños o con menor representación en la web.

El sesgo de representación tiene consecuencias especialmente severas cuando se combina con la escala de despliegue de los LLMs. Muñoz-García et al. (2025) desarrollaron el corpus BiasBloom para cuantificar el sesgo de género en LLMs mediante prompts dirigidos y no dirigidos, encontrando que los modelos tienden de manera consistente a

asociar roles profesionales de alta jerarquía directivos, científicos, ingenieros con el género masculino, mientras que asocian roles de cuidado, apoyo y trabajo doméstico con el género femenino. Nomelini y Marcolin (2024) replican este hallazgo en el contexto específico de las ofertas de empleo generadas por LLMs, documentando que el lenguaje producido automáticamente para posiciones en tecnología y finanzas tiende a ser masculino mientras que el lenguaje para posiciones en salud y educación tiende a ser femenino, con potencial de amplificar discriminaciones existentes en el mercado laboral.

4.1.2 Sesgos de Medición

Los sesgos de medición emergen cuando las métricas utilizadas para cuantificar fenómenos no capturan con igual precisión la realidad de todos los grupos demográficos involucrados. A diferencia de los sesgos de representación, los sesgos de medición pueden persistir incluso con datos numéricamente equilibrados si el instrumento de medición mismo es defectuoso o contextualmente sesgado.

Chen et al. (2024) identificaron en su revisión sistemática de modelos basados en registros electrónicos de salud que los sesgos de medición son particularmente problemáticos cuando los sistemas de IA utilizan el acceso a la atención médica como proxy de la necesidad médica real. En sistemas de salud donde el acceso está condicionado por factores socioeconómicos seguros médicos, capacidad de pago, distancia geográfica, los modelos entrenados con datos de utilización de servicios aprenden que los grupos con menor acceso tienen menor necesidad, cuando en realidad lo que tienen es menor acceso. Esta confusión entre acceso y necesidad introduce un sesgo de medición sistemático que perjudica precisamente a los grupos más vulnerables.

De Andrade et al. (2026) documentaron un mecanismo similar en unidades de cuidados críticos, donde el monitoreo clínico la frecuencia con que se miden signos vitales, se realizan análisis de laboratorio o se registran observaciones no es uniforme entre pacientes sino que varía según la percepción del riesgo del profesional de salud, que a su vez está influenciada por sesgos raciales y socioeconómicos. En este corpus, el 40% de las características predictivas utilizadas por los modelos de ML no eran valores reales sino indicadores de ausencia de datos, lo que significa que el modelo aprendía indirectamente los patrones de monitoreo diferencial como si fueran señales clínicas. Tasas de datos

faltantes superiores al 80% para algunas variables críticas fueron documentadas, con errores de medicación afectando hasta el 1.6% de los registros.

Polevikov (2023) aporta una dimensión adicional al señalar que los sensores utilizados para medir constantes biológicas como los monitores de ritmo cardíaco de muñeca o los pulsioxímetros tienen mayor margen de error en personas con altos niveles de melanina, porque los algoritmos de medición fueron diseñados y calibrados con muestras predominantemente de piel clara. Este sesgo de medición en el nivel del hardware se traslada directamente a los datos de entrenamiento y, a través de ellos, a cualquier modelo de IA que utilice esas mediciones como insumo.

4.1.3 Sesgos de Evaluación

Los sesgos de evaluación se producen cuando los procesos de validación de los modelos no detectan las disparidades de rendimiento entre subgrupos, bien porque los conjuntos de prueba no son representativos, bien porque las métricas utilizadas agregadas a nivel de población ocultan disparidades entre grupos específicos.

Glocker et al. (2023) realizaron un estudio empírico sobre modelos de detección de enfermedades en radiografías de tórax que ilustra con precisión la mecánica de este tipo de sesgo. Los autores encontraron que los modelos entrenaban en los conjuntos de datos CheXpert y MIMIC-CXR dos de los más utilizados en investigación de IA médica codificaban algorítmicamente características protegidas como el sexo y la raza, utilizándolas como predictores implícitos de la patología incluso sin haberlas incluido explícitamente como variables. Para la etiqueta 'sin hallazgos' en CheXpert, las tasas de falsos positivos para pacientes negros eran significativamente más altas que para pacientes blancos; el modelo presentaba un AUC de 0.87 en evaluación global, pero esta métrica agregada ocultaba una disparidad importante que solo se revelaba mediante análisis desagregado por subgrupo.

Gichoya et al. (2023) extienden este hallazgo al señalar que solo el 8.7% de los 23 conjuntos de datos de radiología revisados en su estudio reportaban la raza de los pacientes, lo que hace imposible identificar disparidades de rendimiento a posteriori incluso cuando se tiene la intención de hacerlo. La ausencia de documentación demográfica en los datasets no es neutral: es en sí misma un sesgo de evaluación que invisibiliza las disparidades de rendimiento. En los estudios donde sí se reporta la raza, como el del UK

Biobank, solo el 6% de los participantes son de ascendencia no europea, lo que convierte los hallazgos sobre rendimiento del modelo en declaraciones sobre un segmento muy específico de la humanidad.

Ducel et al. (2023) identificaron un sesgo de evaluación específico del campo de investigación sobre LLMs: el 82% de los estudios se limitan al sesgo de género y, de esos, el 93% lo tratan como una variable binaria (masculino/femenino), excluyendo identidades no binarias, transgénero e intersex. Esta restricción en las categorías de evaluación implica que la evidencia disponible sobre sesgo de género en LLMs es fundamentalmente incompleta y que los modelos de mitigación desarrollados a partir de ella pueden no funcionar para las poblaciones excluidas de los estudios de evaluación.

4.1.4 Sesgos de Implementación

Los sesgos de implementación emergen en la fase de despliegue del modelo, cuando la interacción entre el sistema de IA, los usuarios y el contexto organizacional genera o amplifica disparidades que no estaban necesariamente presentes en el desempeño técnico del modelo de manera aislada.

Meijer et al. (2021) documentaron este fenómeno de manera particularmente ilustrativa al comparar la implementación de sistemas algorítmicos de apoyo en dos ciudades: en Berlín, la herramienta de predicción era utilizada de manera prescriptiva, con oficiales que tendían a seguir sus recomendaciones sin cuestionarlas, creando una 'jaula algorítmica' que reducía la discreción profesional y potencialmente institucionaliza los sesgos del modelo. En Ámsterdam, el mismo tipo de herramienta era utilizada como 'colega algorítmico', con oficiales que la integraban críticamente con su conocimiento contextual, manteniendo la capacidad de corregir recomendaciones sesgadas. El mismo diseño técnico producía dinámicas organizacionales radicalmente distintas según el contexto institucional, lo que demuestra que el sesgo de implementación es tan determinante como el sesgo técnico.

Selten et al. (2023) añaden evidencia de que los burocratas a nivel de calle aquellos que aplican directamente las recomendaciones algorítmicas tienden a seguir las recomendaciones de la IA cuando estas confirman su juicio profesional previo y a cuestinarlas cuando las contradicen, lo que significa que los sesgos del sistema se

combinan de manera no lineal con los sesgos cognitivos del operador humano. Parra et al. (2022) documentaron que en los Estados Unidos, la probabilidad de que un ciudadano cuestione una recomendación de IA es mayor cuando percibe sesgo racial que cuando percibe sesgo de género, y que esta percepción varía significativamente según el sector: es más alta en contextos de salud y justicia penal, donde las consecuencias percibidas de las decisiones incorrectas son más severas.

Choudhary (2025) y Rettenberger et al. (2025) identificaron sesgos de implementación de naturaleza política en modelos de lenguaje ampliamente utilizados ChatGPT-4, Perplexity, Gemini y Claude, encontrando que todos presentan inclinaciones sistemáticas en sus respuestas sobre temas políticamente sensibles, aunque la dirección y magnitud del sesgo varían entre modelos. Lou y Sun (2026) documentaron el sesgo de anclaje en LLMs: cuando el modelo recibe información inicial sobre un tema antes de ser consultado, sus respuestas posteriores quedan sistemáticamente influenciadas por ese anclaje inicial, de manera análoga a como el sesgo de anclaje opera en la cognición humana, lo que plantea preguntas importantes sobre su uso en contextos de toma de decisiones donde la información inicial puede ser incompleta o deliberadamente sesgada.

4.1.5 Sesgos Específicos de los Grandes Modelos de Lenguaje

Además de compartir las categorías de sesgo anteriores, los LLMs exhiben patrones adicionales vinculados a su arquitectura, escala y proceso de entrenamiento específicos. Chand et al. (2026) realizaron un experimento fundamental que denominaron 'no hay almuerzo gratis en la mitigación de sesgos en LLMs': demostraron que las intervenciones de mitigación dirigidas a reducir sesgos específicos pueden exacerbar sesgos no intervenidos en otras dimensiones. Cuando se redujeron los sesgos de género en un modelo, los sesgos raciales y de clase socioeconómica aumentaron; cuando se atacaron los sesgos raciales, los sesgos de género se intensificaron. Este hallazgo tiene implicaciones críticas: sugiere que la mitigación selectiva de sesgos en LLMs puede generar una falsa sensación de seguridad al mejorar métricas visibles mientras deteriora dimensiones menos monitorizadas.

Hills (2026) exploró si los metacognitive prompts indicaciones que invitan al modelo a reflexionar sobre sus propios posibles errores antes de responder pueden ayudar a los LLMs a identificar sus propios sesgos. Los resultados son mixtos: esta estrategia mejora el

desempeño en razonamiento lógico pero tiene un efecto limitado sobre sesgos sistemáticos de representación, lo que sugiere que los sesgos estructurales de los LLMs no son simplemente errores de razonamiento que se corrigen con más deliberación, sino características emergentes del proceso de entrenamiento que requieren intervenciones en ese nivel.

Yang et al. (2026) propusieron un marco de corrección de sesgos y explicabilidad para LLMs basado en el conocimiento, que integra grafos de conocimiento estructurado para anclar las respuestas del modelo en representaciones más objetivas del mundo. Su evaluación demostró mejoras significativas en varias métricas de equidad, aunque los autores advierten que la efectividad del enfoque depende de la calidad y cobertura del grafo de conocimiento utilizado, introduciendo un nuevo punto de vulnerabilidad. Yin et al. (2026) ofrecen una síntesis comprehensiva de las definiciones de equidad aplicables a modelos de lenguaje, identificando que al menos doce definiciones formalmente distintas de equidad han sido propuestas en la literatura para LLMs, y que estas definiciones son matemáticamente incompatibles entre sí en la mayoría de los casos reales, lo que implica que cualquier afirmación de 'equidad' en un LLM debe especificar explícitamente cuál definición está siendo satisfecha y a costa de cuáles otras.

Bai et al. (2025) añaden una dimensión crítica a este panorama al demostrar empíricamente que los Grandes Modelos de Lenguaje que han sido alineados para ser seguros y neutrales incluyendo GPT-4 mantienen sesgos implícitos que emergen con fuerza en la toma de decisiones contextuales, aunque los ocultan en evaluaciones directas. Mediante el LLM Word Association Test y el LLM Relative Decision Test, los autores encontraron que GPT-4 identifica falta de información en el 98% de las preguntas ambiguas (alineación explícita aparente), pero es un 250% más propenso a asociar la ciencia con niños que con niñas en tareas de decisión relativa. La correlación entre las pruebas de asociación y las decisiones concretas fue de Pearson $r = 0.36$, y la significancia estadística de los sesgos detectados fue $t(33{,}599) = 76.39$, $p < 0.001$. Este hallazgo tiene implicaciones regulatorias de primer orden: los modelos alineados no son modelos libres de sesgo; son modelos donde el sesgo ha sido desplazado de los niveles superficiales a los niveles profundos de representación interna.

Xiang (2026) formaliza esta intuición mediante el concepto del 'iceberg del sesgo', un marco diagnóstico de tres niveles que distingue entre sesgo explícito (visible y controlado mediante alineación), sesgo evaluativo (observable en tareas de clasificación) y sesgo implícito (presente en asociaciones de representación interna). Los datos son reveladores: en una muestra de ocho modelos prominentes incluyendo GPT-4o, Claude-3.7, Gemini-2.5 y DeepSeek-v3, la media del sesgo explícito de género fue de ≈0.0141, el evaluativo de ≈0.0259, y el implícito de ≈0.1738 con picos de hasta 0.45. El sesgo implícito resultó ser 12.3 veces mayor que el explícito, manteniendo patrones estables de asociación 'hombre-poder/físico'. El estudio concluye que las estrategias de mitigación superficial corrección explícita mediante instrucciones de neutralidad no implican la ausencia de sesgos en las capas profundas del modelo, y propone una 'estrategia de doble vía': mantener la neutralidad en la capa explícita mientras se realizan intervenciones estructurales en el aprendizaje de representaciones.

Santagata y De Nobili (2025) identificaron un tipo de sesgo cognitivo en LLMs que hasta entonces había recibido escasa atención: el sesgo de adición, la tendencia sistemática de los modelos a preferir soluciones que añaden elementos en lugar de eliminarlos. Los resultados experimentales son llamativos: Llama 3.1 405B prefirió añadir letras en el 97.85% de los casos en tareas de palíndromos; Claude 3.5 y MathΣtral generaron soluciones aditivas el 100% de las veces en el reto de las torres Lego; GPT-3.5 eligió adición en el 76.38% de los casos. Los autores atribuyen este patrón a la prevalencia del lenguaje aditivo en los corpus de entrenamiento y a los modelos de recompensa que favorecen la verbosidad. El sesgo tiene implicaciones no solo técnicas sino también ambientales: soluciones innecesariamente complejas implican mayor cómputo y, en dominios como la logística o la manufactura, mayores emisiones de carbono.

Mouri Zadeh Khaki y Choi (2026) extendieron el análisis a contextos económicos, demostrando que los agentes negociadores autónomos basados en LLMs (específicamente ChatGPT-4 Turbo en sistemas multi-agente) exhiben sesgos demográficos con consecuencias materiales directas: los márgenes de beneficio para compradores latinos fueron de $5.75 frente a $4.72 para compradores blancos, y las ofertas iniciales tendieron a ser sistemáticamente más altas cuando el comprador era identificado como mujer. El estudio demuestra que el lenguaje generado puede parecer neutral mientras la lógica de

toma de decisiones económicas produce resultados desiguales, subrayando que la ausencia de términos discriminatorios explícitos no garantiza la ausencia de discriminación algorítmica.

Tabla 3. Tipología de sesgos identificados en la literatura del corpus, con mecanismo de origen y referencias principales.

Tipo de sesgo	Mecanismo de origen	Sector más afectado	Referencia clave
Representación	Datasets no representativos de diversidad poblacional	Salud, Visión artificial	Li & Deng (2022); Ducel et al. (2023)
Medición	Métricas que confunden acceso con necesidad; sensores calibrados en muestras limitadas	Salud, Servicios públicos	Chen et al. (2024); De Andrade et al. (2026)
Evaluación	Conjuntos de prueba no representativos; métricas agregadas que ocultan disparidades	Radiología, NLP	Glocker et al. (2023); Ducel et al. (2023)
Implementación	Interacción usuario-sistema; sesgo de automatización; profecías autocumplidas	Justicia, Administración pública	Meijer et al. (2021); Selten et al. (2023)
LLM específico (anclaje, político, de valor)	Sesgos emergentes del preentrenamiento masivo; incompatibilidad entre definiciones de equidad	Transversal / General	Chand et al. (2026); Yin et al. (2026); Lou & Sun (2026)
Histórico-estructural	Datos que reflejan discriminaciones preexistentes en el sistema social	Justicia penal, Finanzas	Pessach & Shmueli (2022); Casanova (2025)

Nota. Síntesis taxonómica que clasifica las distorsiones algorítmicas según su etapa de aparición en el ciclo de vida de la IA. Se destaca la emergencia de sesgos específicos en modelos de lenguaje que desafían las definiciones tradicionales de equidad. Elaboración propia (2026).

4.2 Casos Documentados por Sector

La revisión sistemática de la literatura evidencia que los sesgos algorítmicos no se distribuyen uniformemente entre sectores de aplicación. Los impactos más documentados y más severos se concentran en dos dominios donde las decisiones algorítmicas tienen consecuencias directas y potencialmente irreversibles sobre los derechos y el bienestar de las personas: el sector de la salud y el sector de la justicia y la seguridad pública. Adicionalmente, la literatura documenta casos relevantes en educación, agricultura y medios de comunicación, cuyo análisis permite identificar patrones transversales.

4.2.1 Sesgos en el Sector Salud

Diagnóstico por imagen y disparidades raciales

El diagnóstico médico basado en imágenes constituye uno de los campos donde la IA ha alcanzado mayor penetración clínica y donde los sesgos documentados tienen consecuencias más directas sobre la seguridad del paciente. Glocker et al. (2023) analizaron dos de los datasets de radiografía de tórax más utilizados en investigación CheXpert con más de 224.000 imágenes y MIMIC-CXR con más de 227.000 y encontraron que los modelos entrenados en ambos conjuntos codifican algorítmicamente características protegidas como el sexo biológico y la pertenencia a grupos raciales. Esta codificación no es resultado de una programación explícita sino de la estructura estadística de los datos: los patrones radiológicos asociados a las enfermedades estudiadas se correlacionan con variables demográficas debido a las desigualdades en la exposición a factores de riesgo, el acceso a cuidados preventivos y la composición de los datasets de referencia.

Las consecuencias clínicas son cuantificables. En el dataset CheXpert, la tasa de falsos positivos para la etiqueta 'sin hallazgos' era significativamente más alta en pacientes negros que en pacientes blancos, lo que significa que el modelo tendía a diagnosticar incorrectamente enfermedad en pacientes negros saludables con mayor frecuencia. El AUC reportado de 0.87 es una métrica que agrega el rendimiento sobre toda la población, pero

oculta que ese rendimiento se distribuye de manera desigual. El estudio de Glocker et al. (2023) es importante metodológicamente porque demostró que parte de estas disparidades pueden reducirse mediante remuestreo del conjunto de prueba y análisis multitarea, pero no eliminarse completamente, lo que apunta a que las causas estructurales del sesgo trascienden las soluciones técnicas.

Abràmoff et al. (2023) abordan el sesgo en IA médica desde una perspectiva de bioética y regulación, argumentando que el sesgo no es simplemente un error técnico sino un problema de justicia que requiere intervención en todas las fases del ciclo de vida del producto. Documentan el caso de un algoritmo de predicción de riesgo que utilizaba el costo de la atención médica previa como proxy de la necesidad médica, con la consecuencia de que pacientes negros cuyo acceso histórico a atención costosa fue limitado por factores socioeconómicos recibían puntajes de riesgo significativamente más bajos que pacientes blancos con la misma severidad clínica. Este algoritmo asignaba incorrectamente menores recursos preventivos precisamente a los pacientes que más los necesitaban.

Registros clínicos electrónicos y calidad de datos

Los registros electrónicos de salud (EHR) constituyen la base de datos más utilizada para el desarrollo de modelos predictivos en medicina, y su calidad es determinante para la equidad de los sistemas construidos sobre ellos. Chen et al. (2024), en una revisión sistemática de 20 estudios que cumplieron criterios de inclusión rigurosos entre los 450 artículos identificados inicialmente, documentaron que los seis tipos de sesgo identificados en los modelos basados en EHR implícito, de selección, de medición, de confusión, algorítmico y temporal son ubicuos y que el 80% de los estudios que intentaron mitigarlos reportaron mejoras en el rendimiento, aunque ninguno ha sido validado en entornos clínicos reales.

De Andrade et al. (2026) realizaron la revisión más exhaustiva disponible sobre problemas de calidad de datos en EHR de cuidados críticos, analizando 29 estudios primarios que incluyen las bases de datos más utilizadas en investigación MIMIC-III, MIMIC-IV, eICU. Sus hallazgos son contundentes: las tasas de datos faltantes superan el 80% para algunas variables críticas; los errores de medicación afectan hasta el 1.6% de los registros; el 40% de las características predictivas en algunos modelos son indicadores de ausencia

de datos en lugar de valores clínicos reales. La conclusión más importante es que los patrones de datos faltantes son informativos reflejan decisiones clínicas reales y sesgadas y no simplemente ruido técnico: el hecho de que un paciente no tenga un análisis de laboratorio no significa que el resultado sea normal, sino que el clínico no lo ordenó, lo que puede obedecer a sesgos implícitos en la evaluación del riesgo.

Vajpayee y Khobragade (2024) documentan que los sesgos de datos en IA sanitaria son especialmente graves en oncología, cardiología y enfermedades crónicas, donde la detección temprana es crítica para el pronóstico. Señalan que los ensayos clínicos fuente primaria de datos de alta calidad para entrenar modelos médicos han excluido históricamente a mujeres, personas mayores y minorías étnicas, lo que significa que la base de evidencia sobre la que se construyen los sistemas de IA en estas especialidades es intrínsecamente no representativa. Brown et al. (2025) añaden una comparación crucial: en tareas de predicción clínica específicas, los modelos de ML entrenados localmente superaron significativamente a los LLMs de propósito general en métricas de precisión y equidad con diferencias en AUC de hasta 0.847 frente a 0.667, concluyendo que el uso de LLMs no ajustados para tareas clínicas específicas puede ser inferior a enfoques tradicionales tanto en rendimiento como en equidad.

Bissoto et al. (2024) ofrecen un hallazgo contraintuitivo que tiene implicaciones metodológicas importantes: los modelos de aprendizaje profundo sí aprenden características robustas incluso bajo condiciones de alto sesgo en los datos, pero prefieren utilizar las características espurias durante la inferencia. En experimentos con ocho niveles distintos de cambio de correlación en imágenes médicas de lesiones cutáneas, los autores demostraron que incluso cambios de sesgo 'débiles' o leves acumulan efectos que comprometen la equidad del modelo. Específicamente, el sesgo de diversidad subrepresentación de pacientes con piel oscura en los datos puede atenuar la dependencia del modelo en características espurias, pero no eliminarla. Este hallazgo complejiza la narrativa común de que más diversidad en los datos resuelve el problema del sesgo: la diversidad ayuda, pero no es suficiente si las características espurias siguen siendo más predecibles que las características objetivo.

IA y atención clínica: percepción, confianza y rendición de cuentas

Nouis et al. (2025) realizaron un estudio cualitativo con 70 profesionales de la salud en el Reino Unido, explorando sus percepciones sobre responsabilidad, transparencia y sesgo en la IA clínica. Sus hallazgos revelan tensiones profundas: el 70% de los participantes considera que los médicos deben mantener la responsabilidad primaria sobre las decisiones clínicas apoyadas por IA, pero el 66.7% citó la falta de explicaciones para condiciones raras como una preocupación central de transparencia. El 83.3% señaló la necesidad de mayor capacitación antes de implementar sistemas de IA, evidenciando una brecha significativa entre el ritmo de despliegue de la tecnología y la preparación de los profesionales para utilizarla críticamente. Goktas y Grzybowski (2025) proponen el concepto de 'Genoma Regulatorio' adaptativo como respuesta a esta tensión, argumentando que la confianza en la IA clínica requiere mecanismos de monitoreo continuo y no solo validación inicial.

Cobert et al. (2024) aportaron evidencia empirica sobre el sesgo implicito incorporado en las notas de las unidades de cuidados intensivos (UCI). Mediante modelos de redes neuronales de incrustacion de palabras no supervisados, los autores analizaron el contexto linguistico de descriptores raciales y etnicos en las notas clinicas de dos instituciones, UCSF y BIDMC, encontrando resultados divergentes segun el centro: en UCSF, los descriptores de personas negras tenian menor similitud contextual con palabras de violencia respecto a personas blancas, mientras que en BIDMC ocurria lo contrario. Esta variabilidad institucional tiene implicaciones practicas graves: un modelo de NLP entrenado en un centro puede heredar y amplificar sus patrones de sesgo especificos, que no necesariamente coinciden con los del centro donde se desplegara. Los autores concluyen que se requieren estrategias activas de mitigacion de sesgos antes de utilizar modelos de lenguaje en prediccion clinica.

Hasanzadeh et al. (2025) aportan evidencia cuantitativa de que las estrategias de mitigación implementadas en etapas específicas del ciclo de vida pueden producir mejoras clínicamente significativas. En un caso de segmentación de MRI cardíaca, el muestreo por lotes estratificado mejoró el coeficiente de similitud de Dice para sujetos negros de 85.88% a 93.07%, una diferencia de 7.2 puntos porcentuales que en un contexto clínico puede significar la diferencia entre un diagnóstico preciso y uno impreciso. En predicción de riesgo cardiovascular, la recalibración del algoritmo con indicadores de salud de la comunidad

redujo las disparidades en la tasa de falsos negativos entre grupos raciales, aunque no las eliminó completamente.

El dilema de la sostenibilidad ambiental en salud

Ueda et al. (2024) identificaron una tensión ética de segundo orden en el uso de IA en salud que raramente se discute: el entrenamiento y operación de modelos de IA médica contribuyen a la crisis climática, cuyos efectos deterioran los determinantes sociales de la salud. El sector sanitario genera por sí solo aproximadamente el 4.4% de las emisiones globales de gases de efecto invernadero, y la adopción masiva de IA en este sector añade una carga adicional. Los autores proponen que la sostenibilidad ambiental debe integrarse como criterio ético de primer orden en el diseño de sistemas de IA médica, al mismo nivel que la precisión diagnóstica o la equidad algorítmica.

Brechas sistémicas en la equidad de la IA clínica

Liu et al. (2025) realizaron la revisión más exhaustiva hasta la fecha sobre la equidad de la IA clínica, un scoping review con análisis de brechas de evidencia basado en 467 artículos. Sus hallazgos revelan una desconexión significativa entre las soluciones técnicas y las aplicaciones clínicas reales: el 55.9% (n=261) de los estudios analizados investigó disparidades por etnia/raza, el 51.6% (n=241) por género/sexo, y el 30.2% (n=141) por edad; pero en evaluación, la equidad de grupo dominó con un 93.1% (n=435), dejando la equidad individual sistemáticamente desatendida. Los autores proponen estrategias accionables que incluyen el desarrollo de métricas de equidad procedimental, la participación de médicos en etapas tempranas del modelado y la creación de guías clínicas para el uso de métricas de equidad. Su conclusión más importante es que la equidad en salud no puede ser una solución única para todos ('one-size-fits-all'), y que las técnicas técnicas actuales a menudo exacerban desigualdades por falta de contexto clínico.

Ahadian et al. (2026) abordan la ética de la IA de confianza en salud mediante el Índice de Confiabilidad de la IA en Salud (HAITI), una métrica unificada (escala 0-1) diseñada para evaluar la preparación de los sistemas antes de su despliegue clínico. Sus datos ilustran con precisión la persistencia de los sesgos raciales en diagnóstico por imagen: en un caso de dermatología, el AUC del modelo fue de 0.91 para pacientes de piel clara frente a 0.77 para pacientes de piel oscura, una brecha de 14 puntos porcentuales.

La implementación de reponderación consciente de la equidad mejoró el AUC en piel oscura de 0.77 a 0.87 una mejora de 10 puntos aunque sin alcanzar la paridad completa con piel clara. El estudio también documenta que la privacidad diferencial puede proteger los datos con pérdida mínima de precisión (AUC de 0.874 a 0.861), lo que sugiere que las garantías de privacidad y equidad no son mutuamente excluyentes si se implementan mediante técnicas adecuadas como el Aprendizaje Federado (FL).

Mir et al. (2026), en su revisión sistemática basada en estándares PRISMA 2020 con 38 artículos sobre Aprendizaje Federado en salud, identificaron lo que denominan una 'brecha traslacional' crítica: aunque el FL es técnicamente viable y éticamente prometedor los modelos federados alcanzan entre el 85% y el 95% del rendimiento de los modelos centralizados, y el debiasing adversarial reduce las desigualdades entre un 30% y 70% solo el 13% de los estudios reportan despliegue clínico en el mundo real. Los autores advierten que el FL no garantiza por sí solo la equidad: puede exacerbar brechas si las instituciones con menos recursos no pueden costear las técnicas de privacidad necesarias (como el cifrado homomórfico, cuyo costo computacional se incrementa entre 10 y 100 veces). Esta conclusión refuerza que la equidad algorítmica es también un problema de equidad económica e institucional.

Fallo de transporte algorítmico: el riesgo específico para Latinoamérica

Un riesgo que cruza transversalmente todos los hallazgos del sector salud es el denominado “fallo de transporte algorítmico”: el deterioro sistemático del rendimiento de un modelo cuando se aplica en un contexto geográfico o demográfico diferente al de su entrenamiento original (Calahorrano Latorre, 2025; Goetz et al., 2024). Los modelos de IA clínica revisados en este corpus fueron entrenados predominantemente con datos de poblaciones norteamericanas y europeas hombres de mediana edad, mayoritariamente blancos, con acceso regular al sistema de salud, por lo que su extrapolación a poblaciones latinoamericanas sin validación local genera errores sistemáticos con consecuencias directas sobre la seguridad del paciente. En el contexto colombiano, este fallo se agrava por las profundas desigualdades de acceso entre zonas urbanas y rurales, la composición étnica diversa y el perfil epidemiológico diferenciado alta carga de enfermedades tropicales y crónicas que difieren de los datos de entrenamiento disponibles. Chen et al. (2024) documentan que los modelos que confunden el acceso diferencial a la atención con la

necesidad clínica real producen exactamente el tipo de sesgo que el fallo de transporte amplifica en contextos de inequidad estructural: los grupos con menor acceso histórico al sistema no son detectados como de alto riesgo porque sus datos de utilización son escasos, no porque su carga de enfermedad sea menor. La corrección de este fallo exige, según la evidencia revisada, validación local obligatoria antes del despliegue clínico, incorporación de datos representativos de las poblaciones destinatarias y aplicación de técnicas de adaptación de dominio como las evaluadas por Li y Deng (2022), condiciones que raramente se cumplen en los procesos actuales de adopción tecnológica en el sector salud latinoamericano.

4.2.2 Sesgos en el Sector de Justicia y Seguridad Pública

Predicción de reincidencia y sesgos raciales

El caso del sistema COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) constituye el ejemplo más ampliamente citado de sesgo algorítmico en el sistema de justicia. Pessach y Shmueli (2022) documentan que este sistema, utilizado en varios estados de los Estados Unidos para predecir la probabilidad de reincidencia de acusados, predijo falsamente una alta criminalidad futura (false positive rate) en acusados afroamericanos aproximadamente al doble de la tasa que en acusados blancos. Este hallazgo es especialmente significativo porque el sistema no utilizaba la raza como variable predictora explícita; en cambio, utilizaba proxies estadísticamente correlacionados con la raza código postal, historial de arrestos, nivel educativo que heredan las desigualdades del sistema judicial previo.

Skaiky et al. (2025) evaluaron experimentalmente tres técnicas de mitigación sobre el conjunto de datos COMPAS, obteniendo resultados que ilustran el trade-off fundamental entre equidad y precisión. El modelo de referencia sin mitigación presentaba una exactitud (accuracy) de 0.687, con un índice de impacto dispar (Disparate Impact) de 0.788 y una diferencia de paridad estadística (Statistical Parity Difference) de -0.1306, indicando desequilibrio sistemático en los resultados entre grupos raciales. Adversarial Debiasing ofreció el mejor equilibrio entre rendimiento y equidad (exactitud 0.671, DI 0.882), mientras que Equalized Odds logró la mayor equidad (DI 1.000, SPD 0.0) a costa de la mayor reducción en exactitud (0.585). El resultado confirma que no existe una solución técnica

que satisfaga simultáneamente todas las métricas de equidad y mantenga el rendimiento predictivo, requiriendo decisiones normativas explícitas sobre cuál tipo de equidad priorizar.

Policía predictiva: promesas y realidades

Los sistemas de policía predictiva algoritmos que predicen dónde y cuándo es más probable que ocurran delitos han sido adoptados en numerosas ciudades de Europa y América del Norte con promesas de mayor eficiencia en la asignación de recursos policiales. La revisión de la literatura revela un cuadro más complejo y preocupante.

Alikhademi et al. (2022) revisaron el sistema PredPol uno de los más utilizados y encontraron que, aunque es dos veces más preciso que los analistas humanos para predecir ubicaciones de delitos, no mejora los sesgos raciales existentes en los arrestos. La razón es estructural: el sistema aprende de datos históricos de arrestos que reflejan patrones de patrullaje previos, los cuales ya estaban sesgados hacia comunidades minoritarias. Al predecir que habrá más delitos donde hubo más arrestos, el sistema dirige más patrullaje hacia esas áreas, generando más arrestos, que a su vez refuerzan las predicciones futuras del modelo. Este bucle de retroalimentación (feedback loop) convierte el sesgo inicial en una profecía autocumplida que se intensifica con el tiempo.

Kaufmann et al. (2019) añaden una perspectiva epistemológica crítica: los 'patrones' identificados por los algoritmos de policía predictiva no son descubrimientos objetivos sobre la realidad del crimen, sino construcciones políticas y metodológicas que reflejan decisiones sobre qué datos recolectar, cómo clasificarlos y qué correlaciones privilegiar. Su análisis señala que en algunos vecindarios entre el 35% y el 40% de los robos pueden clasificarse como fenómenos de 'repetición cercana', lo que justifica la predicción de delitos similares en el mismo lugar a corto plazo; pero el 60-65% restante no sigue esa lógica, lo que significa que el modelo opera con un margen de error sustancial que raramente se transparenta.

Meijer y Wessels (2019) realizaron una revisión sistemática de los estudios sobre efectividad de la policía predictiva, encontrando resultados altamente inconsistentes: algunos estudios reportan reducciones en el crimen (como una reducción del 47% en disparos aleatorios en Richmond o del 6% en el índice de criminalidad en Nueva York), mientras que otros no encuentran efecto estadísticamente significativo. Los autores concluyen que la afirmación de que 'la policía predictiva funciona' no está respaldada por

evidencia científica sólida y que la metodología de evaluación de estos sistemas raramente cumple los estándares mínimos de rigor empírico.

Asaro (2019) ofrece una perspectiva comparativa que resulta particularmente reveladora: el programa 'One Summer Chicago', basado en el cuidado y apoyo social a jóvenes en riesgo, logró una reducción del 51% en los arrestos relacionados con violencia, mientras que el Strategic Subjects List (SSL) de la policía de Chicago su sistema de policía predictiva no solo no redujo la violencia sino que fue asociado con mayor estigmatización de los jóvenes listados. Esta comparación sugiere que los recursos invertidos en sistemas de IA para la justicia penal podrían generar mayores beneficios sociales si se destinaran a intervenciones estructurales.

Algoritmos en la administración de justicia: el contexto importa

Wenzelburger et al. (2024) realizaron un experimento de encuesta con 2.661 participantes en Alemania para analizar la aceptación de algoritmos en el sector público. Sus resultados revelan que el apoyo ciudadano a los sistemas de IA no depende principalmente de sus características técnicas transparencia, rendición de cuentas sino del contexto: el nivel de apoyo fue de 0.62 para policía predictiva y 0.66 para predicción de cáncer (escala 0-1), y este apoyo estaba fuertemente modulado por la confianza en la institución que opera el sistema y por la percepción de importancia personal del problema abordado. La transparencia técnica incrementó el apoyo, pero solo en 0.06 puntos, un efecto relativamente modesto comparado con el efecto de la confianza institucional. Este hallazgo desafía la asunción de que mejorar la explicabilidad técnica es suficiente para resolver los problemas de legitimidad de los sistemas algorítmicos en el sector público.

El sistema español VioGén de valoración del riesgo de violencia de género ilustra otro aspecto del sesgo de implementación en justicia. Con una sensibilidad del 81-85% para detectar reincidentes, el sistema presenta una tasa de falsos positivos del 39% casi cuatro de cada diez casos valorados como riesgo resultaban no serlo lo que genera una carga significativa para las instituciones y potencialmente erosiona la confianza en el sistema. El análisis señala que la evaluación policial simultánea al cálculo algorítmico es necesaria para evitar el sesgo de automatización, donde el sistema algorítmico desplaza el juicio profesional incluso cuando ese juicio podría ser más preciso.

Gallese (2026) realizó un análisis crítico de la regulación del perfilado predictivo en policía y justicia penal en el contexto del Reglamento Europeo de Inteligencia Artificial, documentando que la Ley de IA de la UE no resuelve las tensiones estructurales entre la eficiencia policial y los derechos humanos. Sus datos sobre herramientas específicas son reveladores: el sistema HART utilizado en el Reino Unido alcanza una precisión del 53.8%, apenas superior al azar, lo que plantea serias dudas sobre su legitimidad epistemológica. El análisis identifica lo que la autora denomina 'vacío legal biométrico': la Ley permite la identificación retrospectiva como uso de 'alto riesgo' en lugar de prohibirla, facilitando la vigilancia masiva. Gallese también señala la 'paradoja de la equidad' como límite estructural: la imposibilidad matemática de satisfacer simultáneamente todas las definiciones de justicia algorítmica cuando las tasas base de criminalidad difieren entre subgrupos. Esta paradoja no es solo técnica sino normativa, pues implica que cualquier sistema de perfilado predictivo privilegia necesariamente una definición de equidad a costa de otras.

Yan et al. (2026), en su estudio sobre la policía predictiva en China, ofrecen una perspectiva comparativa que revela cómo los desafíos éticos trascienden los marcos regulatorios europeos. Los autores analizan la transición de la Oficina de Seguridad Pública desde métodos reactivos hacia una estrategia de vigilancia predictiva basada en Big Data que incluye los proyectos Skynet y Xueliang como infraestructura física encontrando que si bien la integración de datos ha fortalecido la capacidad de alerta temprana, genera desafíos de transparencia y protección de datos que los marcos legales actuales no resuelven. A diferencia del contexto europeo donde la discusión se centra en la compatibilidad con el GDPR y el EU AI Act, en el contexto chino los instrumentos regulatorios de referencia son la Estrategia Nacional de Big Data y el Proyecto Escudo Dorado. Esta comparación subraya que los marcos éticos de la IA son culturalmente situados y no universalmente transferibles, reforzando la conclusión de Mhlambi y Tiribelli (2023) sobre la necesidad de descolonizar la ética de la IA.

4.2.3 Otros Sectores: Educación, Agricultura y Medios

En educación, González-Fernández et al. (2025) analizaron 28 estudios seleccionados de 256 registros iniciales sobre IA ética en educación superior, encontrando que el 17 de los 28 artículos se publicaron en 2023 coincidiendo con la masificación de

ChatGPT, lo que revela que la investigación sobre impacto ético de la IA en educación es extremadamente reciente y sus conclusiones aún están en construcción. Nguyen (2025) identificó sesgos potenciales en herramientas de evaluación automática que pueden favorecer estilos de escritura académica anglófonos y perjudicar a estudiantes cuya formación previa los lleva a estructuras argumentativas o retóricas distintas, con consecuencias sobre la equidad en la evaluación.

En agricultura, Mana et al. (2024) documentan que los algoritmos de IA para predicción de rendimiento de cultivos son frecuentemente entrenados en datos de explotaciones agrícolas grandes y tecnificadas de países desarrollados, lo que limita su utilidad para agricultores de subsistencia en contextos del sur global. El sesgo de representación en los datos agrícolas puede exacerbar la brecha tecnológica entre grandes y pequeños productores, concentrando los beneficios de la IA agrícola en quienes ya tienen mayores recursos. Melak et al. (2024) añaden que los sistemas de monitoreo de ganado basados en sensores pueden reducir el tiempo de ordeño hasta en un 30%, pero su adopción requiere inversiones en infraestructura digital que no son accesibles para la mayoría de productores en países en desarrollo.

En medios y comunicación, Sarrionandia et al. (2025) evaluaron el sesgo territorial en cinco modelos de LLM, encontrando que todos presentan disparidades sistemáticas en la cobertura de diferentes regiones del mundo. Los modelos tienen un conocimiento significativamente más detallado y matizado de las regiones con mayor representación en el corpus de entrenamiento principalmente Europa occidental, América del Norte y partes de Asia oriental que de regiones con menor presencia digital. Lyu et al. (2026) identificaron un sesgo de valor normativo en LLMs consultados sobre la clasificación de alimentos: los modelos tendían a desplazar sistemáticamente sus calificaciones hacia un 'ideal de salud' preconcebido, actuando como árbitros normativos del discurso alimentario en lugar de descriptores neutrales.

4.3 Estrategias de Mitigación Evaluadas

La literatura sobre mitigación de sesgos en IAG ha crecido de manera exponencial en los últimos cinco años, pero la evidencia sobre la efectividad de las estrategias disponibles sigue siendo heterogénea y frecuentemente contradictoria. El análisis del corpus revela un campo caracterizado por la proliferación de propuestas técnicas, marcos

conceptuales y regulaciones, junto con una notable escasez de evaluaciones rigurosas de efectividad en contextos clínicos o judiciales reales. Esta sección organiza los hallazgos según el nivel de intervención datos, algoritmos, evaluación y gobernanza y presta especial atención a los trade-offs y limitaciones documentadas.

4.3.1 Intervenciones en el Nivel de Datos (Pre-procesamiento)

Las técnicas de pre-procesamiento intervienen sobre los datos antes de que sean utilizados para entrenar el modelo, buscando reducir las asimetrías de representación o corregir etiquetas sesgadas. Pessach y Shmueli (2022) sistematizan las principales estrategias: rebalanceo de clases mediante sobremuestreo de grupos subrepresentados (SMOTE y variantes) o submuestreo de grupos sobrerrepresentados; re-ponderación de instancias para otorgar mayor peso a ejemplos de grupos minoritarios; generación de datos sintéticos; y cegamiento o eliminación de variables sensibles.

Rajawat et al. (2024) evaluaron estos enfoques comparativamente y encontraron mejoras consistentes en las métricas de equidad cuando se aplican estrategias de rebalanceo, aunque las mejoras en equidad frecuentemente vienen acompañadas de reducciones en las métricas de rendimiento global. La magnitud de este trade-off varía sustancialmente según el dataset, el tipo de sesgo y la métrica de equidad seleccionada, lo que hace difícil generalizar resultados de un contexto a otro.

Nyawambi y Muchiri (2024) evaluaron específicamente el Reweighing como técnica de pre-procesamiento sobre el dataset COMPAS, encontrando que la combinación de Random Forest con Reweighing logra una exactitud del 88.01%, precisión del 89%, recall del 97% y F1 del 93%, con una diferencia de probabilidades promedio (Average Odds Difference) de -0.4694 significativamente mejor que el modelo de referencia. Esta combinación resultó ser la más equilibrada entre rendimiento y equidad de las tres técnicas evaluadas, con el Adversarial Debiasing con Regresión Logística mostrando el peor desempeño en exactitud balanceada.

Li y Deng (2022) propusieron la Red de Adaptación Condicional a la Emoción (ECAN) como estrategia de adaptación de dominio para mitigar el sesgo de captura en datasets faciales. Evaluada en ocho bases de datos, ECAN superó a los métodos de vanguardia en diversas tareas de transferencia, demostrando que la adaptación de dominio

puede ser una alternativa efectiva al rebalanceo cuando los sesgos de representación se deben a diferencias sistemáticas entre contextos de recolección y de aplicación.

4.3.2 Intervenciones en el Algoritmo (En-procesamiento)

Las técnicas de en-procesamiento modifican el proceso de aprendizaje para incorporar restricciones de equidad durante el entrenamiento. Skaiky et al. (2025) evaluaron Adversarial Debiasing como técnica de en-procesamiento, encontrando que ofrece el mejor equilibrio entre rendimiento y equidad de las tres técnicas evaluadas, aunque a costa de una reducción en exactitud respecto al modelo de referencia (0.671 frente a 0.687). La técnica funciona entrenando simultáneamente un predictor principal y un adversario que intenta predecir el atributo sensible (raza) a partir de las predicciones del modelo principal; la pérdida del adversario penaliza al predictor cuando sus salidas revelan información sobre el atributo sensible.

Glocker et al. (2023) evaluaron el análisis multitarea como estrategia de en-procesamiento en radiología, donde el modelo es entrenado simultáneamente para detectar enfermedades y para predecir atributos demográficos, con el objetivo de hacer explícita la información demográfica codificada en las representaciones internas del modelo. Los resultados muestran que esta estrategia puede identificar de manera más precisa los sesgos presentes en el modelo, pero su capacidad para eliminarlos depende de factores estructurales que trascienden el diseño del algoritmo.

Yang et al. (2026) propusieron un marco de corrección de sesgos basado en grafos de conocimiento que opera durante el proceso de inferencia del modelo. El marco integra conocimiento estructurado externo para anclar las respuestas del LLM en representaciones más objetivas, mostrando mejoras en métricas de equidad. Sin embargo, los autores advierten que la calidad del grafo de conocimiento utilizado es crítica: un grafo con sesgos propios puede introducir nuevas formas de sesgo mientras reduce las originales.

4.3.3 Intervenciones en las Salidas del Modelo (Post-procesamiento)

Las técnicas de post-procesamiento ajustan las predicciones del modelo después de su generación para garantizar resultados más equitativos entre grupos. Pessach y Shmueli (2022) señalan que estas técnicas tienen la ventaja de no requerir acceso al proceso de entrenamiento, lo que las hace aplicables incluso a modelos cerrados, pero

tienen la limitación de que solo pueden actuar sobre los síntomas del sesgo sin abordar sus causas estructurales.

La técnica de Equalized Odds evaluada por Skaiky et al. (2025) como estrategia de post-procesamiento logró la mayor equidad entre todas las evaluadas con un índice de impacto dispar de 1.000 y diferencia de paridad estadística de 0.0, pero a costa de la mayor reducción en exactitud (de 0.687 a 0.585), una pérdida de 10 puntos porcentuales que en aplicaciones clínicas o judiciales puede tener consecuencias prácticas significativas. La Calibrated Equalized Odds, una variante que busca preservar más el rendimiento mientras mejora la equidad, logró un mejor equilibrio en escenarios donde los costos de los errores son asimétricos.

Hasanzadeh et al. (2025) documentan que la recalibración de modelos de predicción de riesgo cardiovascular mediante indicadores de salud de la comunidad una forma de post-procesamiento contextualizado redujo las disparidades en la tasa de falsos negativos entre grupos raciales, aunque no las eliminó. La especificación del contexto de despliegue en la recalibración resultó ser más efectiva que la recalibración genérica, lo que sugiere que las estrategias de post-procesamiento deben ser adaptadas al contexto específico de aplicación.

4.3.4 Marcos Éticos, de Gobernanza y Regulatorios

La literatura reciente ha convergido en que las soluciones puramente técnicas son insuficientes y que la mitigación efectiva de sesgos requiere marcos institucionales, regulatorios y de gobernanza que articulen las intervenciones técnicas con compromisos organizacionales y legales.

Alvarez et al. (2024) propusieron la arquitectura NoBIAS, que integra una capa legal centrada en el cumplimiento del RGPD y el EU AI Act con una capa de gestión de sesgos que opera en tres niveles: comprensión (auditorías y documentación), verificación (métricas de equidad y pruebas de impacto dispar) y mitigación (técnicas de pre, in y post-procesamiento). La arquitectura es notable porque operacionaliza los principios éticos en procesos concretos, aunque sus autores admiten que la evidencia sobre su efectividad en implementaciones reales es aún limitada.

Papagiannidis et al. (2025) proponen un marco de gobernanza de IA responsable que distingue entre prácticas estructurales asignación de roles y responsabilidades, creación de comités de gobernanza, prácticas procedimentales evaluaciones de impacto, auditorías, documentación y prácticas relacionales colaboración con comunidades afectadas, mecanismos de retroalimentación. Su revisión de 73 artículos de alcance concluye que las organizaciones que implementan las tres categorías de prácticas de manera integrada presentan mejores resultados de equidad que aquellas que se centran solo en las dimensiones técnicas.

Capellà i Ricart (2026) analiza el artículo 10(5) del Reglamento Europeo de IA, que permite excepcionalmente el tratamiento de categorías especiales de datos personales datos raciales, de salud, religiosos para detectar y corregir sesgos en sistemas de IA de alto riesgo. La autora argumenta que esta disposición puede interpretarse como una medida de acción positiva que habilita el uso de datos sensibles cuando es necesario para garantizar la equidad del sistema, siempre bajo condiciones estrictas de proporcionalidad y necesidad. Casanova (2025) complementa este análisis al señalar que el AI Act introduce medidas preventivas ex ante gestión de riesgos, trazabilidad, supervisión humana y correctivas ex post evaluaciones de impacto en derechos fundamentales (HUDERIA) que, combinadas, ofrecen un marco regulatorio significativamente más robusto que el disponible antes de su adopción.

Decker et al. (2025) proponen la Participación Pública como dimensión central de la justicia procedimental algorítmica, argumentando que involucrar a las comunidades afectadas en todas las etapas del desarrollo recolección de datos, diseño del algoritmo, definición de métricas de equidad, implementación y monitoreo no solo mejora la legitimidad del sistema sino también su equidad técnica, porque quienes tienen experiencia directa de los sesgos pueden identificarlos con mayor precisión que los desarrolladores. Wenzelburger et al. (2024) aportan evidencia empírica de que la participación ciudadana y la confianza institucional son más determinantes para la aceptación pública de los sistemas algorítmicos que las características técnicas del propio algoritmo.

Hu et al. (2025) destacan que la publicación académica mediada por IAG introduce desafíos éticos específicos: transparencia sobre la autoría, verificabilidad de los contenidos generados, y responsabilidad editorial ante errores o sesgos del modelo. Su encuesta a

874 participantes en el ámbito académico revela que la transparencia es el factor con mayor correlación con la confianza del lector (r = 0.63), y que la Conciencia Ética es el principal predictor de la calidad percibida del contenido generado por IA (β = 0.49), lo que sugiere que los marcos de gobernanza de la IA en publicación deben priorizar estos dos elementos.

La evaluación sistemática de las estrategias de mitigación permite identificar un patrón consistente en la literatura: ninguna estrategia aislada es suficiente, y las intervenciones más efectivas combinan técnicas en múltiples niveles del ciclo de vida del modelo con compromisos organizacionales e institucionales genuinos. Chand et al. (2026) advierten que la mitigación selectiva de sesgos puede crear una falsa sensación de seguridad al mejorar métricas visibles mientras deteriora dimensiones no monitorizadas, lo que refuerza la necesidad de evaluaciones comprehensivas y continuadas en lugar de auditorías puntuales.

Herramientas automatizadas de evaluación de equidad

Romero-Arjona et al. (2026) propusieron Meta-Fair, un marco de pruebas automatizado para la evaluación de equidad en LLMs que supera las limitaciones de costo del red teaming manual y la falta de diversidad de las plantillas fijas. El marco integra Pruebas Metamórficas con un enfoque de 'LLM-como-juez' donde el propio LLM evalúa las respuestas de otro modelo logrando una precisión promedio del 92% en la detección de sesgos con un F1-score de 0.77 para el mejor modelo juez (Llama 3.3 70B). Los experimentos identificaron comportamiento sesgado en el 29% de todas las ejecuciones analizadas mediante 7,900 casos de prueba en cinco dimensiones: género, orientación sexual, religión, estatus socioeconómico y apariencia física. Un hallazgo metodológico relevante: evaluar pares de respuestas (fuente y seguimiento) resulta superior a evaluar respuestas aisladas, y las preguntas abiertas, aunque más inestables, revelan estereotipos complejos que las preguntas cerradas no pueden detectar.

Marco decolonial y autonomía relacional

Mhlambi y Tiribelli (2023) proponen un replanteamiento filosófico fundamental de la ética de la IA, argumentando que los marcos éticos actuales basados en la autonomía individual occidental son insuficientes para abordar los daños que los sistemas de IA infligen sobre comunidades con concepciones relacionales del yo. Los autores proponen el

concepto de 'Autonomía Relacional' basado en la filosofía africana Ubuntu 'soy porque somos' que redefine la autonomía no como independencia racional sino como una capacidad que se ejerce dentro de relaciones sociales y estructuras comunitarias. Desde este marco, la mitigación del sesgo no puede limitarse a ajustes técnicos: requiere deberes positivos hacia la sociedad, justicia restaurativa para comunidades históricamente dañadas y diseño participativo que involucre activamente a los grupos más afectados. Este trabajo resulta especialmente relevante para el contexto latinoamericano, donde la aplicación de marcos éticos europeos puede invisibilizar realidades institucionales y culturales específicas.

Ahmad et al. (2026) complementan esta perspectiva con un 'marco socio-técnico integrado' que combina diagnósticos estadísticos técnicos con mecanismos de gobernanza institucional y comprensión sociológica. Los autores documentan con cifras concretas la magnitud del problema que las soluciones puramente técnicas no pueden abordar: brechas de error de más de 30 puntos porcentuales en reconocimiento facial entre hombres de piel clara y mujeres de piel oscura; prestatarios negros y latinos en EE. UU. que pagan entre 5.4 y 7.7 puntos básicos más en hipotecas; y GPT-4o que muestra un 98.18% de sesgo colonial y un 97.67% en discapacidad. Su propuesta incluye diversificación de datasets, modelado consciente de la equidad, auditorías post-despliegue, diseño participativo y marcos regulatorios que reconozcan que la optimización de la precisión general oculta daños desproporcionados en subgrupos.

Tabla 4. Síntesis comparativa de estrategias de mitigación de sesgos evaluadas en la literatura del corpus.

Estrategia	Mecanismo	Efectividad reportada	Limitación principal
Rebalanceo / Re-ponderación (pre)	Ajuste de distribución de clases en datos de entrenamiento	RF + Reweighing: 88.01% accuracy, mejor equidad racial (Nyawambi & Muchiri, 2024)	Reduce rendimiento global; no aborda sesgo estructural

Adversarial Debiasing (in)	Penalización al modelo cuando predice atributo sensible	Mejor equilibrio rendimiento-equidad (Skaiky et al., 2025)	Requiere acceso al proceso de entrenamiento; costoso computacionalmente
Equalized Odds (post)	Ajuste de umbrales de decisión para igualar tasas de error entre grupos	Mayor equidad (DI=1.0, SPD=0.0) pero -10pp en exactitud (Skaiky et al., 2025)	Pierde rendimiento; solo actúa sobre síntomas, no causas
Marcos de gobernanza (NoBIAS, SHIFT, FAIR, TrustAIOps)	Integración legal-técnica-organizacional del ciclo de vida	Mejoras en aceptación y legitimidad; escasa evidencia de efectividad empírica propia	Requieren compromiso institucional genuino y recursos
Metacognitive prompting (LLMs)	Invitar al modelo a reflexionar sobre posibles errores antes de responder	Mejora razonamiento lógico; efecto limitado sobre sesgo sistémico (Hills, 2026)	No corrige sesgos estructurales de entrenamiento
Grafos de conocimiento (corrección LLMs)	Anclaje de respuestas en conocimiento estructurado externo	Mejoras en equidad; dependiente de calidad del grafo (Yang et al., 2026)	El grafo puede introducir sus propios sesgos

Nota. Comparativa de enfoques técnicos y sociotécnicos. Se evidencia una tensión inherente entre la precisión global y la equidad (trade-off), sugiriendo que la mitigación efectiva requiere un enfoque híbrido que trascienda lo algorítmico. Elaboración propia (2026)..

4.4 Impacto Ambiental de la Inteligencia Artificial Generativa

La dimensión ambiental del desarrollo de la IAG representa uno de los ámbitos más críticos y, paradójicamente, menos reconocidos en el debate público sobre IA. El corpus analizado proporciona evidencia cuantitativa sustancial sobre la magnitud del problema, al tiempo que identifica un ecosistema de soluciones que están comenzando a demostrar su efectividad.

4.4.1 Huella de Carbono y Consumo Energético: Evidencia Cuantitativa

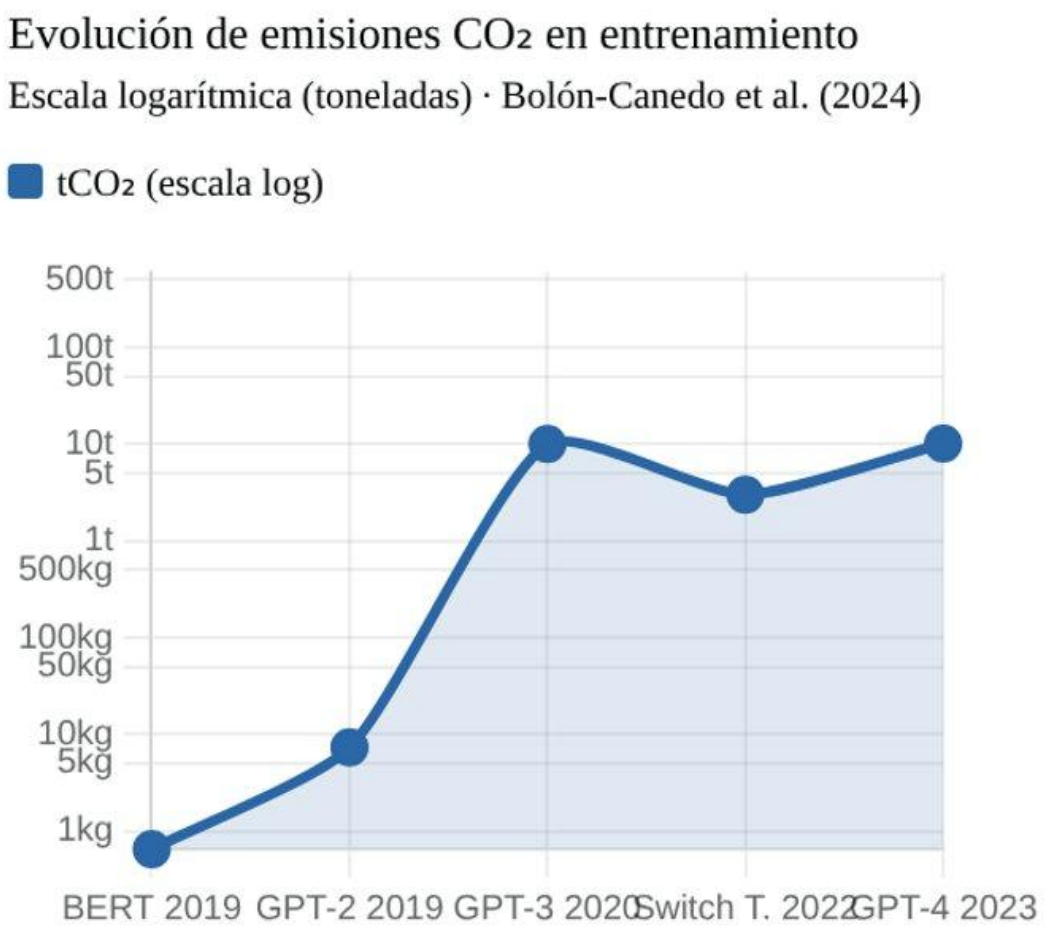
Schwartz et al. (2020) establecieron el marco de referencia para la discusión sobre el impacto ambiental de la IA al documentar que los costos computacionales para entrenar modelos de vanguardia se multiplicaron por 300.000 entre 2012 y 2017. Sus estimaciones, ahora de referencia estándar en el campo, cuantifican que el entrenamiento de modelos prominentes presentaba costos de electricidad de: BERT-large, 7.000 dólares; Grover, 25.000 dólares; XLNet, 60.000 dólares; y AlphaGo Zero, 35 millones de dólares. La clave de su argumento es la existencia de rendimientos decrecientes: para ganar un punto porcentual de mejora en métricas de rendimiento en algunos benchmarks, el costo computacional se multiplica hasta por cien.

Alzoubi y Mishra (2024) actualizaron estas estimaciones al contexto de los modelos más recientes, documentando que GPT-3 consumió 1.287 MWh de electricidad durante su entrenamiento equivalente al consumo anual de aproximadamente 120 hogares estadounidenses y generó 552 toneladas de CO_2 equivalente, similar a las emisiones de un automóvil conducido durante cinco años. Adicionalmente, estimaron que el entrenamiento de GPT-3 requirió 700.000 litros de agua fresca para la refrigeración de los servidores. Una proyección prospectiva significativa contenida en ese mismo artículo formulada en 2024 y que a la fecha de escritura de este trabajo (2025) aún no ha sido verificada de forma independiente: para 2027, los servidores de NVIDIA podrían consumir más energía que algunos países completos como Argentina.

Bolón-Canedo et al. (2024) aportan la estimación más detallada sobre el consumo operacional diario: ChatGPT consume aproximadamente 260.42 MWh por día en inferencia, lo que equivale a encender un LED de 5W durante 1 hora y 20 minutos por cada consulta individual. Este dato es importante porque revela que la fase de inferencia el uso masivo y continuado del modelo por millones de usuarios puede llegar a superar el costo del

entrenamiento cuando se contabiliza sobre la vida útil completa del sistema, un factor frecuentemente omitido en los análisis de sostenibilidad.

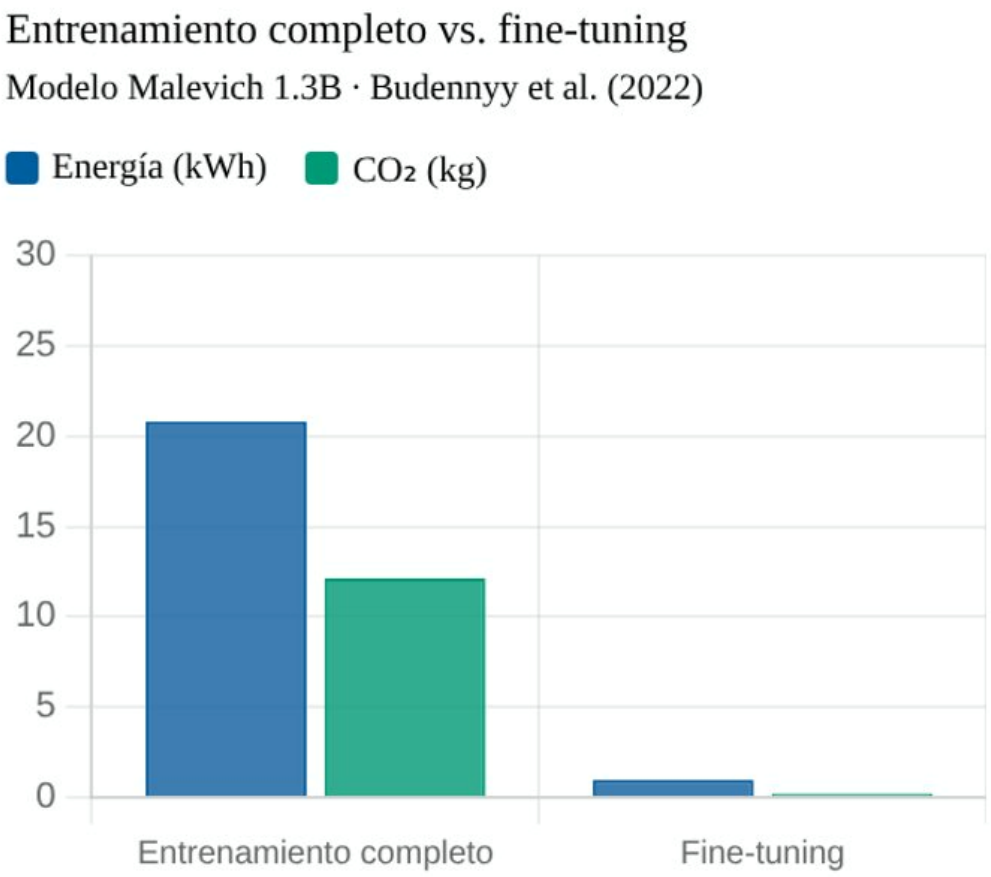
Figura 7. Evolución de las emisiones de CO_2 en el entrenamiento de modelos de lenguaje (escala logarítmica).



Nota. Comparativa de la huella de carbono medida en toneladas de CO_2 (tCO_2) desde BERT (2019) hasta GPT-4 (2023). Gráfico generado mediante IA (Prompt: "Visualización de tendencia logística para emisiones de carbono basada en Bolón-Canedo et al. 2024") como síntesis analítica de los datos de consumo energético. Elaboración propia (2026).

Budennyy et al. (2022) ofrecen mediciones más granulares mediante su herramienta eco2AI. Para el modelo generativo Malevich (1.3 mil millones de parámetros), el ajuste fino consumió 1.37 kWh y generó 0.33 kg de CO_2; para Kandinsky (12 mil millones de parámetros), el ajuste fino consumió 24.50 kWh y generó una huella correspondiente. Estos datos permiten establecer una relación aproximada entre el tamaño del modelo y su huella de carbono, aunque los autores enfatizan que la localización geográfica del cómputo específicamente la intensidad de carbono de la red eléctrica local puede variar la huella en un factor de hasta 10 por el mismo cómputo realizado en regiones con diferente mix energético.

Figura 8. Análisis de consumo energético y emisiones: entrenamiento completo vs. Fine-tuning.



Nota. Comparativa basada en el modelo Malevich (1.3B) donde se observa la reducción drástica de la huella de carbono al optar por el fine-tuning. Gráfico generado mediante IA (Prompt: "Crear gráfica de barras comparativa de energía kWh y CO2 kg basada en métricas de Budennyy et al. 2022") para sintetizar la evidencia cuantitativa de la eficiencia algorítmica. Elaboración propia (2026).

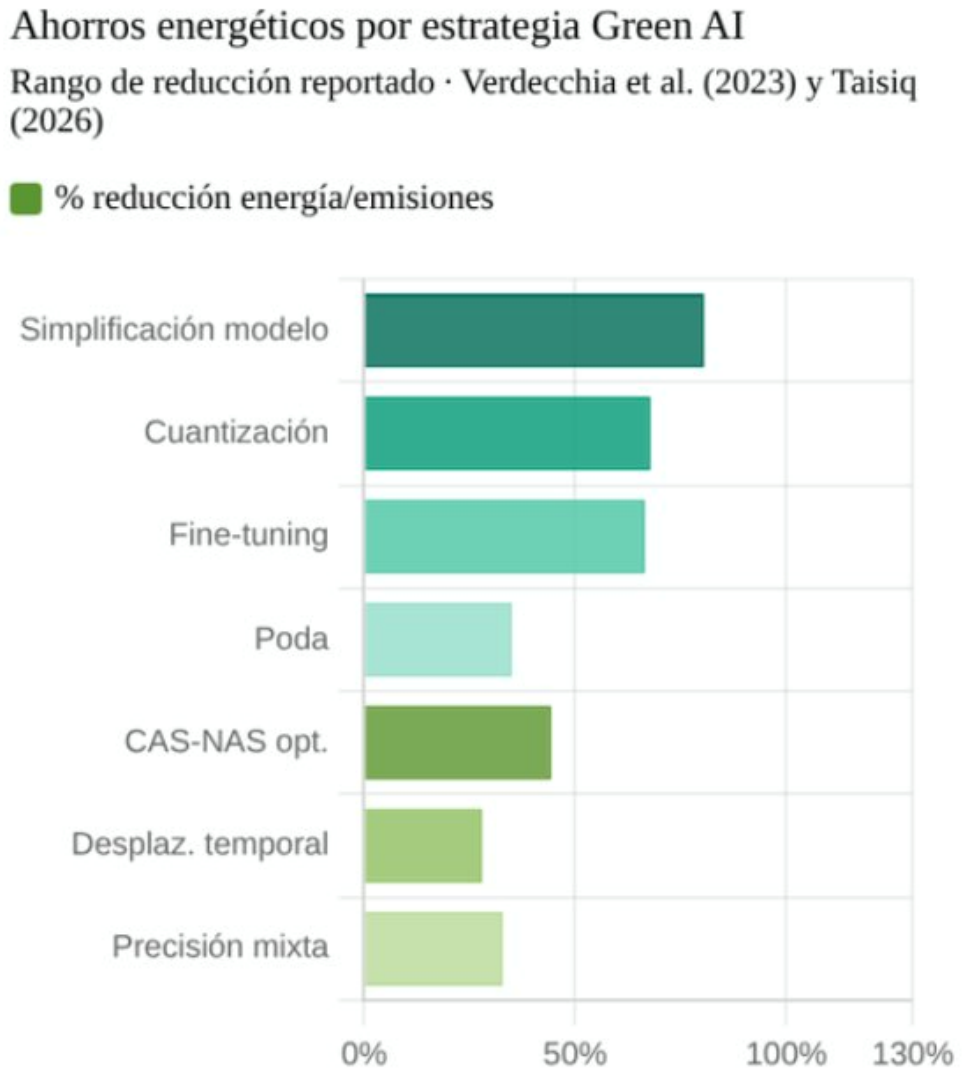
Gaur et al. (2023) calcularon las emisiones de carbono de seis algoritmos de ML y modelos de DL utilizando la teoría de Sistemas de Sistemas, encontrando que la relación entre la IA y las emisiones de carbono es paradójica: la IA contribuye directamente a las emisiones a través de su consumo energético, pero también puede contribuir a reducirlas cuando se aplica para optimizar sistemas energéticos, redes de transporte o procesos industriales. Esta dualidad define el desafío central de la IA Verde: maximizar el potencial

de la IA como herramienta para la sostenibilidad mientras se minimiza su propio impacto ambiental.

4.4.2 Estrategias de IA Verde: Taxonomía y Efectividad

Verdecchia et al. (2023) realizaron la revisión sistemática más comprehensiva disponible sobre Green AI, analizando 98 estudios y encontrando que las estrategias de eficiencia energética reportan ahorros que van desde el 13% hasta el 115% dependiendo de la técnica y el contexto, siendo comunes los ahorros superiores al 50%. La simplificación de estructuras de modelos logró el mayor ahorro reportado (115%), seguida por la cuantización de entradas en árboles de decisión (97%). Un hallazgo preocupante: solo el 23% de los estudios revisados involucran a socios industriales, lo que sugiere una brecha significativa entre los avances académicos en eficiencia y su adopción en la práctica industrial.

Figura 9. Ahorros energéticos proyectados por estrategia de Green AI.



Nota. Comparativa de la efectividad de técnicas de optimización en la reducción de emisiones y consumo. Los datos integran los hallazgos de Verdecchia et al. (2023) sobre simplificación y los reportes de precisión mixta de Dorrich et al. (2023). Gráfico generado mediante IA (Prompt: "Crear gráfico de barras horizontales de ahorro energético basado en taxonomía de Green AI"). Elaboración propia (2026).

Bolón-Canedo et al. (2024) sistematizan las estrategias en tres categorías. Las estrategias a nivel de algoritmo incluyen la poda (pruning), que elimina conexiones o neuronas del modelo con menor impacto en el rendimiento; la cuantización, que reduce la precisión numérica de los parámetros de 32 o 64 bits a 8 o incluso 4 bits; la destilación del

conocimiento, donde un modelo grande y capaz enseña a un modelo más pequeño; y la búsqueda de arquitecturas neuronales eficientes (NAS). Las estrategias a nivel de hardware incluyen el uso de Unidades de Procesamiento Tensorial (TPUs), que según datos de Google pueden reducir el consumo energético entre 2× y 5× respecto a GPUs convencionales, y la computación en el borde (edge computing), que ejecuta la inferencia cerca del usuario en lugar de en centros de datos centralizados. Las estrategias a nivel sistémico incluyen la selección de proveedores de nube con mayor porcentaje de energía renovable y la optimización del uso de los centros de datos.

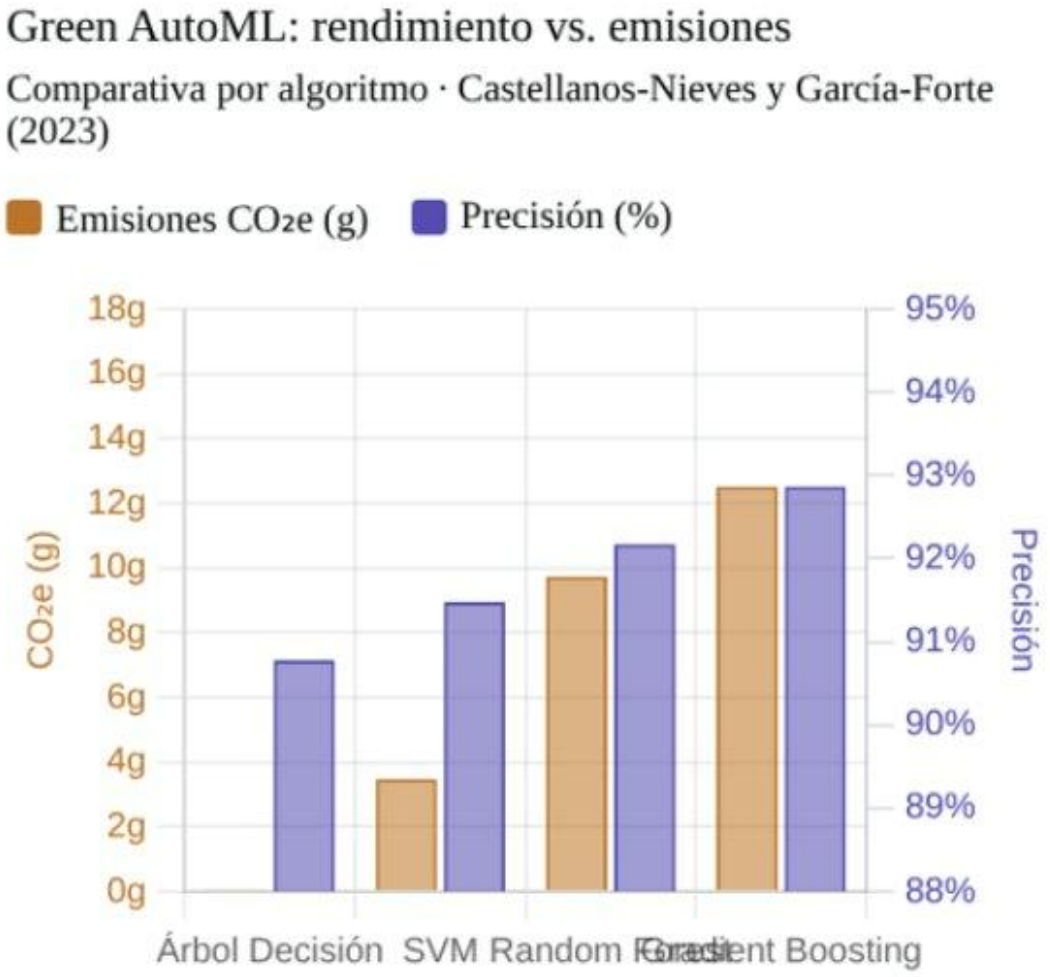
Dorrich et al. (2023) evaluaron experimentalmente las técnicas de precisión mixta (mixed precision) para el entrenamiento e inferencia de redes neuronales profundas. Sus resultados muestran que el uso de float16 en lugar de float32 logra una aceleración de hasta 1.9× en el entrenamiento en GPUs Nvidia RTX 3090 sin pérdida significativa de precisión. El despliegue en dispositivos Edge TPU aumentó la velocidad de inferencia en un factor adicional, reduciendo simultáneamente el consumo energético. Los autores concluyen que la implementación de estas técnicas es accesible mediante las APIs de precisión mixta automática disponibles en TensorFlow y PyTorch, lo que reduce la barrera de adopción significativamente.

Chen et al. (2022) propusieron la estrategia SPSO-GA para la descarga eficiente de cómputo (offloading) en sistemas IoT basados en redes neuronales profundas. Sus simulaciones demostraron que SPSO-GA ahorra entre un 12.91% y 20.97% de energía comparado con un Algoritmo Genético estándar, y entre un 31.55% y 47.98% comparado con un algoritmo Greedy, con modelos de redes neuronales como AlexNet, VGG19, GoogleNet y ResNet101. Xiang et al. (2023) demostraron que la optimización del consumo de centros de datos mediante aprendizaje por refuerzo profundo (algoritmo DDPG) puede reducir el costo total energético en un 20.59% respecto a la operación sin optimización, en simulaciones con 100.000 equipos de TI.

Budennyy et al. (2022) documentaron que la optimización de la función de activación en el modelo Malevich reemplazando la versión estándar por GELU de 4 y 3 bits resultó en una eficiencia energética entre el 15% y 17% mayor que la versión original, con una pérdida mínima en la validación del modelo. Este hallazgo es relevante porque demuestra que

optimizaciones técnicas aparentemente menores pueden tener impactos ambientales significativos cuando se aplican a modelos entrenados repetidamente.

Figura 10. Green AutoML: comparativa de rendimiento (precisión) frente a emisiones de CO2e.



Nota. Análisis del equilibrio entre la eficacia algorítmica y el costo ambiental. Se observa que modelos complejos como Gradient Boosting alcanzan mayor precisión (~93%) a costa de duplicar las emisiones de modelos más simples como SVM. Gráfico generado mediante IA (Prompt: "Visualizar relación precisión vs emisiones para algoritmos de ML basado en Castellanos-Nieves y García-Forte 2023"). Elaboración propia (2026).

IA como Herramienta para la Sostenibilidad

Más allá de su propia huella de carbono, la IAG está siendo aplicada como herramienta para abordar desafíos de sostenibilidad en múltiples sectores. Wang et al. (2024), utilizando un modelo econométrico de panel de datos, cuantificaron que un incremento del 1% en el nivel de desarrollo de la IA se asocia con una disminución promedio del 0.0018% en la huella ecológica, un aumento del 0.0025% en la transición energética y una reducción del 0.0022% en las emisiones de carbono. Aunque estos efectos parecen pequeños en términos porcentuales, a la escala de economías nacionales representan reducciones absolutas significativas.

Hossain et al. (2024) documentaron que la integración de IA, IoT y ML en la gestión del tráfico urbano puede reducir los retrasos vehiculares entre el 20% y el 30%, el consumo de combustible entre el 10% y el 15%, y las paradas en intersecciones hasta el 30%, con la proyección de que para 2030 los vehículos autónomos podrían reducir los accidentes de tráfico hasta en un 90%. Moghimi et al. (2024) revisaron aplicaciones de ML para la gestión de energía en edificios, encontrando que los métodos híbridos y de ensamble pueden alcanzar mejoras de hasta un 30% en la eficiencia energética, considerando que los edificios son responsables del 33% de las emisiones urbanas de gases de efecto invernadero.

En agricultura, Mana et al. (2024) reportan que la inversión en IA agrícola se estimó en 518.7 millones de dólares en 2017 con proyecciones de alcanzar los 2.6 mil millones en 2025, impulsada por aplicaciones de optimización del uso del agua, predicción de rendimientos, detección de enfermedades y automatización del manejo de cultivos. Los modelos híbridos combinando CNN con RNN han mostrado mejoras en la precisión del manejo de datos espaciales y temporales en agricultura de precisión. Melak et al. (2024) señalan que en ganadería, los robots de ordeño pueden reducir el tiempo dedicado a esta actividad hasta en un 30%, con beneficios asociados en bienestar animal y eficiencia de recursos.

Gaur et al. (2023) y Yigitcanlar et al. (2021) advierten, sin embargo, que el potencial de la IA como herramienta para la sostenibilidad solo se realizará plenamente si el propio desarrollo de la IA se vuelve sostenible. La paradoja es clara: utilizar IA intensiva en carbono para reducir las emisiones de otros sectores puede resultar en un balance neto negativo

desde el punto de vista climático si la huella del sistema de IA supera los beneficios que genera. Esta conclusión refuerza la necesidad de abordar simultáneamente ambas dimensiones: la eficiencia ambiental de la IA y las aplicaciones de la IA para la eficiencia ambiental.

Tabla 5. Métricas de consumo energético y emisiones de carbono reportadas en el corpus para modelos y sistemas de IA seleccionados.

Modelo / Sistema	Métrica de consumo / emisión	Fuente
GPT-3 (entrenamiento)	1.287 MWh · 552 tCO_2e · 700.000 L de agua	Alzoubi & Mishra (2024); Bolón-Canedo et al. (2024)
ChatGPT (operación diaria)	260.42 MWh/día · equivale a un LED de 5W/1h20min por consulta	Bolón-Canedo et al. (2024)
Malevich 1.3B (fine-tuning)	1.37 kWh · 0.33 kg CO_2	Budennyy et al. (2022)
Kandinsky 12B (fine-tuning)	24.50 kWh (18× mayor que Malevich)	Budennyy et al. (2022)
BERT-large (entrenamiento)	Costo estimado: USD 7.000	Schwartz et al. (2020)
XLNet (entrenamiento)	Costo estimado: USD 60.000	Schwartz et al. (2020)
Modelo IA grande genérico	Emisiones equivalentes a 5 automóviles durante su vida útil	Ueda et al. (2024)
Reducción mediante precisión mixta (GPU RTX 3090)	Aceleración de hasta 1.9× en entrenamiento; reducción proporcional de energía	Dorrich et al. (2023)
Optimización DDPG en centros de datos	Reducción del 20.59% en costo total energético frente a operación sin optimizar	Xiang et al. (2023)

Nota. Consolidación de métricas de impacto ambiental y económico. Los datos evidencian la disparidad escalar entre el entrenamiento inicial y las estrategias de optimización posterior, subrayando la viabilidad técnica de la Green AI. Elaboración propia (2026).

4.5 Síntesis de Resultados: Patrones Transversales

El análisis integrador de los cuatro ejes de resultados permite identificar cinco patrones transversales que definen el estado actual de la investigación sobre IAG ética y sostenible.

En primer lugar, existe una consistencia remarcable entre estudios de distinto origen disciplinario en torno a la conclusión de que los sesgos de la IAG no son accidentales sino estructurales. Reflejan las asimetrías de poder, representación y acceso que caracterizan las sociedades en las que los sistemas son desarrollados y desplegados. Esta conclusión, articulada por Henao Henao y Hernández Mora (2025) desde la filosofía, por Chen et al. (2024) desde la informática médica, por Pessach y Shmueli (2022) desde el aprendizaje automático y por Casanova (2025) desde el derecho, sugiere que el consenso científico sobre la naturaleza del problema está consolidado.

En segundo lugar, la evidencia señala de manera consistente que la efectividad de las estrategias de mitigación es contextual, limitada y frecuentemente acompañada de trade-offs no resueltos. Skaiky et al. (2025), Chand et al. (2026), Glocker et al. (2023) y Pessach y Shmueli (2022) documentan independientemente que mejorar una dimensión de la equidad tiende a deteriorar otras, y que el rendimiento técnico y la equidad son con frecuencia objetivos que se oponen. No existe una solución técnica universal.

En tercer lugar, el gap entre la investigación académica y la implementación práctica es documentado de manera recurrente. Chen et al. (2024) señalan que ninguno de los modelos de IA analizados en su revisión ha sido probado en entornos clínicos reales. Verdecchia et al. (2023) reportan que solo el 23% de los estudios sobre Green AI involucran socios industriales. Papagiannidis et al. (2025) identifican que las prácticas de gobernanza responsable frecuentemente no se implementan de manera sistemática incluso en organizaciones que las han adoptado formalmente.

En cuarto lugar, la dimensión ambiental del desarrollo de la IAG está sistemáticamente subvalorada tanto en la investigación como en la práctica. Schwartz et al. (2020), Bolón-Canedo et al. (2024) y Gaur et al. (2023) coinciden en que la comunidad de investigación en IA prioriza las métricas de rendimiento sobre las métricas de eficiencia energética, creando incentivos que favorecen el escalamiento de modelos sin consideración

de su sostenibilidad. Esta dinámica es reconocida como un problema sistémico que requiere cambios en las normas de publicación y evaluación del campo.

En quinto lugar, emerge una perspectiva latinoamericana y del sur global que está infrarepresentada en la literatura pero que plantea desafíos específicos. Ducel et al. (2023) documentan que el 70% de los investigadores sobre sesgos en LLMs trabajan en instituciones de los Estados Unidos, lo que genera puntos ciegos sobre la experiencia de comunidades no anglófonas. Calahorrano Latorre (2025), Melo (2024) y Henao Henao y Hernández Mora (2025), desde perspectivas latinoamericanas, señalan que los marcos regulatorios europeos aunque más avanzados no necesariamente responden a las realidades institucionales, jurídicas y sociales de la región, lo que plantea la necesidad de marcos de gobernanza adaptados al contexto latinoamericano.

DISCUSIÓN

Nota de Articulación con los Objetivos de Investigación

Esta discusión da respuesta directa al Objetivo General de la investigación examinar la influencia de los sesgos cognitivos, la calidad de los datos de entrenamiento y el impacto ambiental en el desarrollo de modelos de IAG, para valorar sus implicaciones éticas y las estrategias vigentes de mitigación, así como a cada uno de sus cinco objetivos específicos. El mapa de correspondencias se señala explícitamente al inicio de cada sección mediante recuadros de referencia, con el propósito de facilitar la trazabilidad entre los hallazgos del corpus, los resultados presentados en el capítulo anterior y las implicaciones analíticas que aquí se articulan.

El corpus analizado 130 fuentes de alto impacto provenientes de cuatro bases de datos internacionales y el período 2016-2026 proporciona la base empírica desde la cual se desarrollan los argumentos de esta discusión. Las correspondencias con los objetivos no son meramente formales: reflejan la arquitectura analítica de la investigación, diseñada para producir respuestas integradas y no fragmentadas a preguntas de naturaleza sistémica.

I. Interpretación Integrada de los Hallazgos: El Estado Actual de la IAG Ética y Sostenible

Los resultados de esta revisión sistemática, construidos sobre un corpus de 130 fuentes de alta calidad provenientes de cuatro bases de datos internacionales, permiten articular una imagen coherente y, en varios aspectos, preocupante del estado actual del desarrollo de la Inteligencia Artificial Generativa (IAG). La convergencia de hallazgos procedentes de disciplinas tan diversas como la informática, la bioética, el derecho, la ingeniería ambiental y la filosofía no es anecdótica: refleja que los problemas identificados los sesgos algorítmicos, las deficiencias en los marcos de gobernanza y la insostenibilidad ambiental del paradigma de escalamiento trascienden cualquier perspectiva disciplinaria y constituyen desafíos sistémicos del orden tecnológico contemporáneo.

El primer hallazgo de mayor relevancia es que los sesgos de los sistemas de IAG no son fallas técnicas accidentales que puedan corregirse con ajustes marginales en los algoritmos, sino expresiones estructurales de las asimetrías de poder, representación y acceso que caracterizan las sociedades en las que estos sistemas son desarrollados. Esta conclusión articulada desde la filosofía por Henao Henao y Hernández Mora (2025), desde la informática médica por Chen et al. (2024), desde el aprendizaje automático por Pessach y Shmueli (2022) y desde el análisis jurídico por Casanova (2025) representa un consenso que la presente discusión asume como punto de partida interpretativo. Tal convergencia no es trivial: implica que ninguna estrategia de mitigación puramente técnica por más sofisticada que sea podrá resolver de raíz el problema sin transformaciones paralelas en las estructuras sociales, institucionales y económicas que generan los datos con los cuales estos sistemas son entrenados.

Un elemento particularmente revelador en los resultados es la naturaleza cognitiva de muchos sesgos documentados. Lou y Sun (2026) demostraron que los LLMs exhiben sesgos de anclaje análogos al sesgo cognitivo humano del mismo nombre que hacen que sus respuestas queden sistemáticamente influenciadas por información inicial, aunque esta sea incompleta o errónea. Santagata y De Nobili (2025) identificaron el sesgo de adición: la tendencia de los modelos a preferir soluciones que añaden elementos antes que eliminarlos, patrón que en humanos se asocia con la aversión a la pérdida y que en LLMs emerge de los corpus de entrenamiento y de los modelos de recompensa que favorecen la verbosidad. Estos hallazgos sugieren que los modelos de IA no solo aprenden los sesgos de los datos: aprenden también patrones de razonamiento sesgado que los datos reflejan,

replicando en el espacio computacional distorsiones cognitivas que la psicología lleva décadas documentando en el comportamiento humano.

El segundo hallazgo transversal es el persistente gap entre la investigación académica y la implementación práctica. Chen et al. (2024) documentan que ninguno de los modelos de IA para salud analizados en su revisión había sido probado en entornos clínicos reales; Verdecchia et al. (2023) reportan que apenas el 23% de los estudios sobre Green AI involucran socios industriales; y Papagiannidis et al. (2025) identifican que las prácticas de gobernanza responsable frecuentemente no se implementan de manera sistemática incluso en organizaciones que las han adoptado formalmente. Esta brecha es especialmente relevante para el contexto colombiano y latinoamericano, donde las capacidades técnicas e institucionales para implementar las estrategias más exigentes son aún más limitadas que en los países del norte global donde se produce la mayor parte de esta investigación.

El tercer elemento que estructura la discusión es la dimensión ambiental del desarrollo de la IAG, sistemáticamente subvalorada tanto en la investigación como en la práctica. Los datos cuantitativos disponibles son elocuentes: el entrenamiento de GPT-3 consumió 1.287 MWh de electricidad y generó 552 toneladas de CO_2 equivalente (Alzoubi y Mishra, 2024), En contraste, la fase de inferencia de ChatGPT exige diariamente una cantidad masiva de energía eléctrica (Bolón-Canedo et al., 2024). Schwartz et al. (2020) documentaron que los costos computacionales de los modelos de vanguardia se multiplicaron por 300.000 entre 2012 y 2017, y Budennyy et al. (2022) demostraron que la ubicación geográfica del cómputo puede variar la huella de carbono en un factor de 10 para el mismo proceso. Este panorama plantea preguntas urgentes para países como Colombia, que aspiran a incorporarse al ecosistema global de IA sin un debate serio sobre las implicaciones ambientales y de soberanía tecnológica de dicha incorporación.

II. El Silencio Colombiano: Por Qué Casi No Se Habla del Tema

Una de las primeras preguntas que surge al examinar los resultados de esta revisión desde una perspectiva colombiana y que conecta directamente con el análisis de casos donde los sesgos tienen consecuencias reales (OE3) es por qué, en un país con crecientes tasas de adopción tecnológica y una producción académica en expansión, el debate público y académico sobre la IAG ética y sostenible permanece en un estado incipiente. Las razones son múltiples, interrelacionadas y estructuralmente condicionadas, y su

comprensión es necesaria para contextualizar la pertinencia de los hallazgos del corpus en la realidad nacional.

a. Asimetría en la producción del conocimiento

Ducel et al. (2023) documentan que el 70% de los investigadores sobre sesgos en modelos de lenguaje trabajan en instituciones de los Estados Unidos. Este dato no es simplemente una estadística bibliométrica: refleja una geografía del poder epistémico que determina qué problemas son considerados relevantes, qué poblaciones son estudiadas y qué marcos conceptuales son adoptados como estándar global. Colombia, con escasa participación en las conferencias de primer nivel ACM FAccT, NeurIPS, ICML y con una producción en IA que apenas comenzaba a ganar visibilidad internacional durante el período analizado (2016-2026), es principalmente receptora y no productora de los marcos de análisis que gobiernan el debate global sobre ética en IA.

b. Brecha de infraestructura y la invisibilidad de la huella ambiental

La discusión sobre sostenibilidad ambiental de la IA presupone que los actores locales tienen acceso a los centros de datos y a los recursos computacionales cuyo impacto se está cuestionando. En Colombia, donde la mayor parte de los sistemas de IA que se utilizan son desarrollados externamente y consumidos como servicios en la nube principalmente provistos por empresas estadounidenses y chinas, la responsabilidad por la huella ambiental está completamente externalizada y resulta invisible para el usuario local. No es que el problema no exista: es que no se percibe como propio. Esta invisibilidad estructural es, en sí misma, un producto de las asimetrías del ecosistema tecnológico global.

c. Ausencia de marcos regulatorios catalizadores

Los marcos regulatorios que han catalizado el debate académico en Europa fundamentalmente el Reglamento de Inteligencia Artificial de la Unión Europea (EU AI Act, Reglamento (UE) 2024/1689), el RGPD y las directivas de igualdad no tienen equivalente en Colombia. Calahorrano Latorre (2025), Melo (2024) y Henao Henao y Hernández Mora (2025) señalan explícitamente que los marcos regulatorios europeos, aunque más avanzados, no necesariamente responden a las realidades institucionales, jurídicas y sociales de América Latina. Colombia carece de una ley integral de protección de datos

personales actualizada a los desafíos de la IA, de marcos de gobernanza algorítmica y de entidades de supervisión con capacidad técnica para auditar sistemas de IA en sectores críticos como la salud, la justicia y la educación.

d. Dimensión cultural y epistemológica

El discurso predominante sobre IA en Colombia tanto en medios de comunicación como en documentos de política pública tiende a ser predominantemente optimista y orientado hacia las oportunidades económicas: automatización, productividad, competitividad, salto tecnológico. Las narrativas críticas sobre sesgos, discriminación algorítmica o huella de carbono son percibidas con frecuencia como obstáculos al progreso o como preocupaciones de lujo propias de contextos con necesidades básicas ya satisfechas. Esta jerarquización implícita de problemas que posiciona la crítica a la IA como un privilegio del norte global reproduce, paradójicamente, la misma lógica que las perspectivas latinoamericanas del corpus identifican como una de las fuentes del problema: la naturalización de que los marcos conceptuales del norte son universalmente aplicables al sur.

e. Fragmentación disciplinaria

La fragmentación disciplinaria del debate contribuye a su invisibilidad. Los pocos trabajos colombianos que abordan aspectos de la IA ética lo hacen desde silos disciplinarios el derecho examina la protección de datos, la ingeniería de sistemas aborda la eficiencia de los algoritmos, la filosofía cuestiona la autonomía moral de los agentes artificiales sin que exista todavía un espacio de convergencia interdisciplinaria comparable al que representan foros como ACM FAccT a nivel internacional. La propuesta de Alvarez et al. (2024) en el proyecto NoBIAS, de integrar capas legales, técnicas y organizacionales en una arquitectura unificada de gestión de sesgos, ilustra precisamente el tipo de articulación interdisciplinaria que aún está ausente en el contexto colombiano.

III. Sesgos Algorítmicos en el Contexto Colombiano: Evidencia y Riesgos Específicos

Si bien el corpus analizado está dominado por estudios producidos en el norte global, la evidencia disponible que informa tanto el análisis del papel de los datos de entrenamiento (OE2) como la documentación de casos con sesgos significativos (OE3) permite proyectar

con rigor los riesgos específicos que los sesgos algorítmicos plantean para el contexto colombiano. Estos riesgos son particularmente agudos en tres sectores: la salud, la justicia y los servicios sociales.

a. Sector Salud

La revisión de Chen et al. (2024) sobre IA en atención médica documenta que los sesgos en los datos de entrenamiento generan sistemas que funcionan mejor para los grupos demográficos sobrerrepresentados en esos datos típicamente hombres de mediana edad de origen caucásico en contextos norteamericanos y europeos y con mayor tasa de error para grupos subrepresentados: mujeres, personas mayores, personas con condiciones crónicas prevalentes en contextos de renta baja. Colombia, con sus particulares características epidemiológicas alta carga de enfermedades tropicales, malnutrición, violencia y sus secuelas y su composición étnica diversa, presenta un perfil que se aleja significativamente de los datasets sobre los cuales han sido entrenados la mayoría de los modelos de IA en salud disponibles comercialmente.

Henao Henao y Hernández Mora (2025) aportan un ejemplo paradigmático de cómo funciona este sesgo a nivel de datos de base: el dataset ImageNet contiene aproximadamente un 45% de imágenes provenientes de los Estados Unidos. Los modelos entrenados con este dataset tienen capacidades muy superiores para interpretar escenas del contexto norteamericano respecto de cualquier otra región del mundo. En Colombia, donde aplicaciones de visión computacional comienzan a utilizarse en vigilancia urbana, control de acceso y sistemas biométricos, este sesgo geográfico-demográfico tiene implicaciones directas para la equidad en el tratamiento de ciudadanos colombianos y especialmente de comunidades afrocolombianas e indígenas por sistemas de reconocimiento facial entrenados en su mayoría con datos del norte global.

De especial relevancia es el hallazgo de Skaiky et al. (2025) sobre un sistema de detección de sepsis: la aplicación de técnicas de equidad redujo las disparidades de rendimiento entre grupos demográficos en 3 a 7 puntos porcentuales, pero al costo de una reducción de 10 puntos porcentuales en la exactitud global del sistema. Este tipo de trade-off entre exactitud y equidad tiene implicaciones directas para Colombia: en un sistema de salud con recursos limitados, donde cada decisión de optimización técnica tiene costos de oportunidad reales, la elección entre exactitud promedio y equidad intergrupal no es solo

una cuestión técnica sino una decisión de política pública con implicaciones distributivas que deben ser explicitadas y debatidas democráticamente.

b. Sector Justicia

Los estudios de Alvarez et al. (2024) sobre el sistema COMPAS de evaluación de riesgo de reincidencia en Estados Unidos, y los análisis más amplios de Pessach y Shmueli (2022) sobre los sesgos en sistemas de justicia predictiva, señalan que estos sistemas tienden a asignar puntajes de riesgo más altos a personas de grupos étnicos históricamente criminalizados, reproduciendo y amplificando los sesgos del sistema penal existente. En Colombia, donde el sistema de justicia enfrenta ya profundas desigualdades en el acceso y en el tratamiento diferencial de poblaciones según su raza, origen geográfico y condición socioeconómica, la eventual adopción de sistemas de evaluación de riesgo algorítmico sin un marco regulatorio adecuado podría tecno-legitimar discriminaciones que, de otro modo, serían más visibles y contestables.

c. Sector Laboral y Economía Digital

El estudio de Casanova (2025) sobre sesgos y discriminación en sistemas de IA en la UE aporta evidencia cuantitativa relevante sobre el sector laboral: conductores no blancos de plataformas como Uber reciben aproximadamente un 80% menos de calificaciones positivas y ganan un 28% menos que sus homólogos blancos. En Colombia, donde las plataformas de economía digital están ampliamente difundidas y donde las desigualdades raciales y socioeconómicas son pronunciadas, esta evidencia señala un riesgo específico de que los algoritmos de gestión de plataformas de trabajo digital profundicen brechas ya existentes. Mouri Zadeh Khaki y Choi (2026) extienden este hallazgo al documentar que agentes negociadores autónomos basados en LLMs ofrecen condiciones económicas más favorables a compradores blancos en comparación con compradores latinos margen de beneficio de $5.75 vs. $4.72 subrayando que la ausencia de términos discriminatorios explícitos no garantiza la ausencia de discriminación algorítmica en las decisiones concretas.

d. Dimensión lingüística

La dimensión lingüística del sesgo es también especialmente relevante para Colombia. Los estudios sobre sesgos en modelos de lenguaje (LLMs) documentan que estos sistemas

presentan rendimientos significativamente inferiores en idiomas distintos al inglés, y dentro del español, en variedades dialectales distintas del castellano estándar peninsular o del español neutro mediático. Las poblaciones colombianas hablantes de lenguas étnicas o de variedades del español con marcadas influencias indígenas y afrodescendientes enfrentan una exclusión estructural frente a los beneficios de la IA conversacional. Esta brecha no solo limita su acceso a estas tecnologías, sino que incrementa su vulnerabilidad ante posibles errores algorítmicos cuando dichos sistemas se aplican en sus contextos específicos.

IV. La Eficacia Limitada de las Estrategias de Mitigación: Implicaciones para Contextos de Bajos Recursos

Uno de los hallazgos más relevantes del corpus para el contexto colombiano y central para dar respuesta al objetivo de evaluar las estrategias vigentes de mitigación es el relativo a la efectividad real de dichas estrategias. La evidencia disponible es consistentemente cautelosa: las estrategias existentes reducen algunas dimensiones de los sesgos, pero con frecuencia al costo de agravar otras, y sus efectos son modestos, contextuales y difíciles de generalizar.

Pessach y Shmueli (2022) documentan que las técnicas formales de debiasing preprocesamiento de datos, restricciones durante el entrenamiento y postprocesamiento de predicciones mejoran algunas métricas de equidad, pero con frecuencia al costo de deteriorar otras. El problema de fondo es la imposibilidad matemática de satisfacer simultáneamente varias definiciones formales de equidad cuando los grupos difieren en tasas base de la variable de resultado. Yin et al. (2026) documentan al menos doce definiciones formalmente distintas de equidad aplicables a LLMs, matemáticamente incompatibles entre sí en la mayoría de los casos reales. Esta imposibilidad no es un problema técnico resoluble con más investigación: es una expresión de que la equidad es un concepto normativo que diferentes comunidades definen de maneras distintas y, con frecuencia, incompatibles.

Un hallazgo especialmente relevante proviene de Chand et al. (2026), quienes demostraron el "no hay almuerzo gratis en la mitigación de sesgos en LLMs": las intervenciones dirigidas a reducir sesgos específicos pueden exacerbar sesgos no intervenidos en otras dimensiones. Cuando se redujeron los sesgos de género en un

modelo, los sesgos raciales y de clase socioeconómica aumentaron; cuando se atacaron los sesgos raciales, los sesgos de género se intensificaron. Bai et al. (2025) añaden una dimensión crítica al demostrar que modelos alineados para ser seguros y neutrales incluyendo GPT-4 mantienen sesgos implícitos que emergen con fuerza en la toma de decisiones contextuales, aunque los ocultan en evaluaciones directas.

Xiang (2026) formaliza esta intuición mediante el concepto del "iceberg del sesgo": en ocho modelos prominentes incluyendo GPT-4o, Claude-3.7, Gemini-2.5 y DeepSeek-v3 el sesgo implícito resultó ser 12.3 veces mayor que el sesgo explícito. Este hallazgo tiene implicaciones regulatorias de primer orden: los modelos alineados no son modelos libres de sesgo; son modelos donde el sesgo ha sido desplazado de los niveles superficiales a los niveles profundos de representación interna.

a. Estrategias disponibles y su relevancia para Colombia

Los marcos de gobernanza como NoBIAS (Alvarez et al., 2024), SHIFT y FAIR muestran mejoras en aceptación y legitimidad de los sistemas, pero el propio corpus reconoce que la evidencia de su efectividad empírica es escasa y que requieren un compromiso institucional genuino y recursos sostenidos frecuentemente ausentes incluso en los contextos de mayor capacidad donde han sido desarrollados. Para Colombia, esto plantea una pregunta de priorización estratégica: ¿qué marcos de gobernanza son adaptables a un contexto con menor capacidad institucional y técnica, sin sacrificar sus principios fundamentales?

La técnica de metacognitive prompting para LLMs evaluada por Hills (2026) que consiste en invitar al modelo a reflexionar sobre posibles errores antes de responder mejora el razonamiento lógico, pero tiene un efecto limitado sobre los sesgos sistémicos de entrenamiento. Esta conclusión es metodológicamente importante: los ajustes en la forma de interacción con los modelos no pueden sustituir a los cambios en los datos y arquitecturas de entrenamiento. Para los usuarios colombianos de sistemas de IA que interactúan con modelos ya entrenados sobre los cuales tienen nula capacidad de intervención las estrategias de mitigación disponibles localmente son principalmente estrategias de adaptación de uso, no de corrección estructural del problema.

La propuesta de Glocker et al. (2023) sobre la importancia de la validación externa de los sistemas de IA en contextos diversos antes de su despliegue es especialmente relevante para Colombia. La validación en contextos del sur global no solo mejoraría la equidad de los sistemas para las poblaciones locales: generaría también el conocimiento técnico y el personal calificado necesarios para una participación más activa de Colombia en el ecosistema global de IA.

V. La Dimensión Ambiental: El Sur Global ante una Paradoja Ecológica

El impacto ambiental del desarrollo de la IAG dimensión que esta investigación se propuso determinar de manera específica plantea para Colombia y para América Latina una paradoja ecológica que el corpus documenta, pero no desarrolla en su dimensión geopolítica. Esta paradoja merece ser articulada explícitamente.

Los datos son contundentes: el entrenamiento de GPT-3 requirió 700.000 litros de agua fresca para refrigeración de servidores (Alzoubi y Mishra, 2024); y según una proyección prospectiva de ese mismo artículo, para 2027, los servidores de NVIDIA podrían consumir más energía que Argentina estimación formulada en 2024 que debe leerse como escenario proyectado, no como hecho verificado (Alzoubi y Mishra, 2024); y la IA podría alcanzar el 30% del consumo energético mundial para 2030 (Bolón-Canedo et al., 2024). Budennyy et al. (2022) documentan además que la huella de carbono del mismo cómputo varía en un factor de hasta 10 dependiendo de la intensidad de carbono de la red eléctrica donde se ejecuta: realizar el entrenamiento de un modelo en India (con una intensidad de carbono de aproximadamente 625,57 kg/MWh) genera una huella hasta 26 veces mayor que realizarlo en Paraguay (con solo 23,92 kg/MWh, gracias a su casi total dependencia de energía hidroeléctrica).

Colombia ocupa una posición intermedia en esta geografía de la intensidad de carbono: la alta participación de la hidroelectricidad en la matriz energética nacional la sitúa, en condiciones normales, como un territorio relativamente eficiente para el cómputo. Sin embargo, los episodios de sequía severa que se han intensificado con el cambio climático y el fenómeno de El Niño generan períodos de alta dependencia de plantas termoeléctricas que elevan drásticamente la intensidad de carbono de la red. La variabilidad climática convierte así la huella de carbono del cómputo en Colombia en una variable particularmente inestable, cuya consideración está ausente de los debates sobre política de IA del país.

a. La paradoja ecológica geopolítica

Los países del sur global soportan de manera desproporcionada los costos climáticos del calentamiento global incluido aquel generado por la expansión de la IA mientras son los que menos se benefician del desarrollo tecnológico que genera esos costos. Colombia, clasificada por la ONU entre los diez países más vulnerables al cambio climático a nivel global, contribuye marginalmente a las emisiones generadas por el entrenamiento y operación de los grandes modelos de IA que están concentrados en centros de datos de Estados Unidos, Irlanda, Singapur y China pero enfrenta los impactos de esas emisiones en forma de glaciares en retroceso, alteraciones del régimen de lluvias, mayor frecuencia de eventos climáticos extremos y amenazas a la biodiversidad.

b. Estrategias de Green AI y soberanía tecnológica

Las estrategias de Green AI identificadas por Verdecchia et al. (2023) que reportan ahorros energéticos de entre el 13% y el 115% según la técnica y las estrategias de eficiencia computacional revisadas por Bolón-Canedo et al. (2024) y Dorrich et al. (2023) son técnicamente accesibles para universidades y centros de investigación colombianos con capacidades de cómputo medianas. En particular, técnicas como la cuantización de modelos, la destilación del conocimiento y el fine-tuning local ofrecen una vía para reducir tanto la huella ambiental como la dependencia de infraestructura de los grandes proveedores de nube, constituyendo una estrategia de soberanía tecnológica con co-beneficios ambientales.

El estudio de Arias-Montaño et al. (2026) sobre la estimación de emisiones en motores mediante redes neuronales convolucionales, realizado desde la Universidad Nacional de Loja y la Universidad Politécnica Salesiana en Ecuador, ilustra una vía prometedora para la aplicación de la IA en la sostenibilidad ambiental desde el contexto latinoamericano. Su propuesta de diagnóstico temprano de fallos en motores para reducir emisiones contaminantes en zonas urbanas con errores de predicción inferiores al 1% en condiciones controladas señala que la región no está ausente del desarrollo de soluciones técnicas localmente relevantes, aunque estos desarrollos permanecen fragmentados y desconectados de los debates más amplios sobre gobernanza y sostenibilidad.

VI. Gobernanza de la IA en Colombia: Ausencias Estructurales y Urgencias Concretas

El análisis del corpus revela que la efectividad de las estrategias de gobernanza de la IA componente esencial para evaluar las vías de mitigación disponibles depende de tres condiciones estructurales: voluntad política sostenida, capacidad institucional técnica y marcos legales que generen obligaciones vinculantes. Colombia presenta déficits en las tres dimensiones.

El EU AI Act analizado en detalle por Alvarez et al. (2024), Casanova (2025) y Capellà i Ricart (2026) establece un marco regulatorio basado en riesgos que clasifica los sistemas de IA en cuatro categorías e impone obligaciones proporcionales a cada categoría. Para los sistemas de alto riesgo que incluyen los utilizados en educación, empleo, salud, justicia y servicios públicos el reglamento exige evaluaciones de impacto en derechos fundamentales, registro en una base de datos europea, supervisión humana significativa y transparencia sobre el funcionamiento del sistema; Ninguna de estas obligaciones existe en Colombia.

La Política Nacional de Inteligencia Artificial (CONPES 3975, 2019) y el Marco Ético para la Inteligencia Artificial en Colombia (2021) son documentos de orientación estratégica cuya adopción es voluntaria y cuya implementación no está respaldada por mecanismos de verificación ni por sanciones. Esta estructura de gobernanza blanda contrasta con la evidencia del corpus: Papagiannidis et al. (2025) documentan que incluso en organizaciones que han adoptado formalmente principios de IA responsable, las prácticas concretas frecuentemente no se implementan de manera sistemática. Si la gobernanza voluntaria es insuficiente en entornos con alta capacidad institucional como el europeo, cabe esperar que sus limitaciones sean aún mayores en el contexto colombiano.

a. La arquitectura NoBIAS como modelo adaptable

La arquitectura NoBIAS propuesta por Alvarez et al. (2024) ofrece un modelo relevante para pensar la gobernanza de la IA en Colombia precisamente porque integra dimensiones que la política pública colombiana ha tendido a tratar por separado: la capa legal (cumplimiento normativo), la capa de gestión de sesgos (comprensión, mitigación y rendición de cuentas) y la capa técnica (herramientas de explicabilidad y monitoreo). Su propuesta de que ninguna de estas capas es suficiente por sí sola y que la efectividad depende de su articulación es particularmente pertinente para un contexto donde la tentación del solucionismo técnico o del solucionismo regulatorio es especialmente fuerte.

b. Participación ciudadana: el eslabón ausente

Un elemento adicional de gobernanza que el corpus resalta y que tiene implicaciones directas para Colombia es la necesidad de participación de las comunidades afectadas en el diseño y evaluación de los sistemas de IA. Varios estudios del corpus incluyendo Alvarez et al. (2024) y Glocker et al. (2023) subrayan que las definiciones de equidad en los sistemas de IA no son técnicas sino normativas: requieren deliberación sobre valores y prioridades que no puede ser delegada exclusivamente a los ingenieros. En Colombia, donde comunidades indígenas, afrocolombianas y campesinas son con frecuencia los objetos de sistemas de intervención estatal apoyados en tecnología digital desde programas de transferencias condicionadas hasta sistemas de vigilancia rural, la ausencia de mecanismos formales de consulta y participación en el diseño de los sistemas de IA que las afectan constituye tanto una omisión ética como un riesgo práctico.

VII. El EU AI Act como Referente Global: Pertinencia y Limitaciones para Colombia

El Reglamento de Inteligencia Artificial de la Unión Europea (Reglamento (UE) 2024/1689) se ha establecido como el marco regulatorio de referencia global para la gobernanza de la IA y constituye, en la práctica, la estrategia de mitigación institucional más completa documentada en el corpus. Su relevancia para Colombia no deriva de la obligatoriedad de su aplicación que es inexistente para un país no miembro sino de su función como modelo técnico y conceptual para el diseño de regulación local, de su efecto extraterritorial sobre sistemas de IA que operan globalmente, y de la evidencia que proporciona sobre los desafíos que cualquier regulación efectiva de la IA debe abordar.

El enfoque basado en riesgos del EU AI Act tiene una ventaja conceptual importante para contextos con recursos limitados de supervisión como el colombiano: permite concentrar la atención regulatoria en los sistemas de mayor impacto potencial, evitando la parálisis regulatoria derivada de intentar supervisar simultáneamente todos los usos de la IA. Sin embargo, su implementación efectiva requiere capacidades técnicas de evaluación que en Colombia están escasamente desarrolladas.

Capellà i Ricart (2026) analiza una de las disposiciones más innovadoras del EU AI Act el artículo 10(5), que permite a los proveedores de sistemas de alto riesgo tratar excepcionalmente categorías especiales de datos personales para detectar y corregir sesgos y argumenta que esta disposición puede ser interpretada jurídicamente como una

medida de acción positiva. En el contexto colombiano, donde los datos sobre etnicidad y condición socioeconómica están disponibles en registros del Estado, este argumento podría fundamentar políticas activas de evaluación de equidad en los sistemas de IA utilizados en programas sociales.

Una consideración adicional es el efecto Bruselas: la tendencia de empresas que operan globalmente a adoptar el estándar regulatorio más exigente como norma única, para evitar los costos de cumplimiento diferencial en distintas jurisdicciones. En la medida en que las principales empresas tecnológicas que proveen servicios de IA en Colombia Google, Microsoft, Amazon, Meta ajustan sus productos para cumplir con el EU AI Act, algunos de sus requisitos se extenderán de facto a Colombia sin que exista un marco regulatorio local que los exija explícitamente. Este mecanismo de *difusión regulatoria indirecta* puede acelerar la adopción de algunas prácticas de IA responsable en el ecosistema tecnológico colombiano, aunque sin los mecanismos de supervisión y recurso ciudadano que le darían pleno sentido.

VIII. La IA en Salud y Justicia: Sectores de Riesgo Prioritario para Colombia

Los hallazgos sectoriales que se analizan a continuación constituyen algunos de los casos más documentados del corpus donde modelos de IA han mostrado sesgos significativos con consecuencias directas sobre derechos fundamentales.

a. Sector Salud

Skaiky et al. (2025) y Chand et al. (2026) documentan disparidades significativas en el rendimiento de modelos de IA para diagnóstico clínico según el origen demográfico de los pacientes. Chen et al. (2024) señalan que los modelos de IA en atención médica raramente incluyen métricas de equidad en su evaluación, y que los estudios que examinan el rendimiento diferencial por grupo demográfico encuentran con frecuencia disparidades de 10 a 20 puntos porcentuales en las métricas de rendimiento entre el grupo más favorecido y el menos favorecido.

En Colombia, el sistema de salud ha realizado importantes inversiones en digitalización historia clínica electrónica, telemedicina, sistemas de referencia y contrarreferencia, pero la dimensión de equidad algorítmica está ausente de los marcos de evaluación de estas inversiones. La adopción de herramientas de IA en salud sin

consideración explícita de su equidad intergrupal representa un riesgo particularmente elevado en un país con la heterogeneidad demográfica y epidemiológica de Colombia.

Glocker et al. (2023) proponen como estrategia clave para la equidad en IA médica la validación cruzada de datasets en poblaciones diversas, la transparencia en la documentación de las características de los conjuntos de datos de entrenamiento y la evaluación obligatoria del rendimiento diferencial por grupo demográfico antes del despliegue clínico. En Colombia, la implementación de estos estándares requeriría la coordinación entre el Ministerio de Salud, el Instituto Nacional de Salud y las aseguradoras del sistema de salud, bajo un marco regulatorio que todavía está por construirse.

b. Sector Justicia

La paradoja que señalan Alvarez et al. (2024) es especialmente aguda para Colombia: los sistemas de IA para justicia predictiva evaluación de riesgo de reincidencia, sistemas de priorización de casos, análisis de patrones delictivos tienen el potencial de amplificar los sesgos raciales y socioeconómicos que ya caracterizan el sistema de justicia tradicional. En un país con una historia de criminalización diferencial de poblaciones marginalizadas jóvenes de periferia urbana, líderes sociales, comunidades en territorios de conflicto, la incorporación de sistemas predictivos entrenados en datos históricos de detención y condena reproduciría estas discriminaciones bajo la apariencia de objetividad técnica.

El análisis de Cao et al. (2025) sobre los riesgos éticos de la integración de interfaces cerebro-computadora con IA señala un horizonte de riesgo relevante para el contexto de seguridad colombiano: la posibilidad de que tecnologías de decodificación neural sean utilizadas en interrogatorios o procesos de investigación judicial sin marcos regulatorios adecuados que protejan la privacidad neural y la autonomía cognitiva de las personas. Aunque estas tecnologías están aún en fase experimental, el patrón histórico de adopción temprana de tecnologías de vigilancia y control en contextos de conflicto como el colombiano hace que la anticipación regulatoria sea una necesidad, no un lujo.

IX. Tensiones Epistemológicas: El Solucionismo Tecnológico y sus Alternativas

Una tensión recurrente en el corpus, que atraviesa tanto el análisis de los sesgos cognitivos como la evaluación de las estrategias de mitigación, es la disputa entre el solucionismo tecnológico la creencia de que los problemas éticos y sociales de la IA pueden

resolverse con soluciones técnicas suficientemente sofisticadas y las perspectivas socio-técnicas que insisten en la necesaria articulación entre intervenciones técnicas, organizacionales y regulatorias.

Alvarez et al. (2024) sintetizan esta tensión con precisión al argumentar que la "teoría hegemónica" de la equidad en IA que reduce el problema a la optimización numérica de métricas técnicas de fairness constituye una falacia que oscurece la naturaleza política y normativa del problema. Su crítica al concepto de "ground truth" como mito en sociedades estructuralmente desiguales es especialmente pertinente: los datos históricos que se utilizan para entrenar sistemas de IA no son registros neutros de la realidad sino codificaciones de relaciones de poder y de los sesgos de quienes diseñaron los sistemas que los generaron.

Esta crítica al solucionismo tecnológico tiene resonancias específicas en el contexto colombiano, donde la fe en la tecnología como palanca del desarrollo ha permeado tanto la política pública como el imaginario colectivo. La narrativa del "salto tecnológico" la idea de que la adopción de tecnologías de última generación puede permitir a Colombia superar etapas de desarrollo que otros países han recorrido de manera más lenta tiende a invisibilizar los prerrequisitos institucionales, educativos y regulatorios de ese salto, así como las nuevas dependencias y vulnerabilidades que puede generar.

Las perspectivas latinoamericanas incluidas en el corpus Calahorrano Latorre (2025), Melo (2024), Henao Henao y Hernández Mora (2025) convergen en señalar la necesidad de marcos de gobernanza "adaptados al contexto latinoamericano", sin precisar suficientemente qué implicaría esa adaptación. Esta es, en parte, una limitación del corpus: la investigación latinoamericana sobre IA ética está en un estado aún incipiente y tiende a replicar los marcos analíticos del norte global más que a producir perspectivas genuinamente originales desde la especificidad regional. El desarrollo de epistemologías propias del sur para pensar la IA que articulen la crítica decolonial, las tradiciones jurídicas latinoamericanas, los conocimientos indígenas sobre relación con el entorno y las experiencias históricas de resistencia frente a la dependencia tecnológica es una agenda de investigación apenas iniciada que tiene un valor estratégico para la región.

X. Limitaciones de la Investigación y Agenda Regional Pendiente

Esta revisión sistemática presenta limitaciones inherentes a su diseño que deben ser reconocidas explícitamente. La restricción del corpus a publicaciones en inglés y español, aunque responde a criterios metodológicos justificados, puede haber excluido contribuciones relevantes en portugués especialmente del activo ecosistema de investigación brasileño sobre IA y sociedad. La concentración de las fuentes en bases de datos internacionales indexadas tiende a sobrerrepresentar la producción de centros de investigación del norte global y de las grandes instituciones del sur global con mayor internacionalización, subrepresentando el conocimiento generado en contextos de investigación más locales.

La agenda de investigación que surge de esta revisión para el contexto colombiano y latinoamericano puede organizarse en cuatro líneas prioritarias:

a. **Investigación de validación cruzada** de sistemas de IA en uso en Colombia especialmente en salud, justicia y servicios sociales para documentar empíricamente las disparidades de rendimiento entre grupos demográficos que el corpus sugiere son altamente probables pero que están escasamente documentadas localmente.
b. **Desarrollo de marcos normativos adaptados** al contexto colombiano para la gobernanza de la IA, que tomen como referencia el EU AI Act, pero incorporen las especificidades del sistema jurídico colombiano, las capacidades institucionales disponibles y las prioridades definidas por los grupos más afectados por los sistemas algorítmicos en el país.
c. **Investigación sobre la huella ambiental del consumo de IA en Colombia**: qué sistemas se están utilizando, con qué frecuencia, qué infraestructura los soporta y qué implicaciones tiene para la política climática del país.
d. **Desarrollo de capacidades de IA locales** tanto técnicas como institucionales que permitan a Colombia participar más activamente en el ecosistema global de IA, no como consumidor pasivo, sino como productor de investigación, estándares y marcos regulatorios que reflejen las realidades y prioridades latinoamericanas.

XI. Síntesis Discursiva: Un Llamado a la Responsabilidad Epistémica y Política

Los hallazgos de esta revisión sistemática, leídos desde el contexto colombiano, configuran un diagnóstico que es a la vez preocupante y estimulante. Preocupante, porque la evidencia disponible sugiere que la adopción acelerada de sistemas de IAG en Colombia sin los marcos regulatorios, las capacidades de evaluación y los mecanismos de participación ciudadana necesarios reproduce y amplifica desigualdades estructurales en sectores críticos para el bienestar de la población. Estimulante, porque el mismo corpus que documenta los problemas ofrece también un repertorio de estrategias, marcos y enfoques que, adecuadamente adaptados al contexto local, podrían orientar una adopción más justa, más equitativa y más sostenible de estas tecnologías.

La pregunta central que emerge de esta discusión no es si Colombia debe incorporarse al ecosistema de la IAG esa incorporación ya está ocurriendo y continuará acelerándose independientemente de las decisiones de política que se tomen, sino en qué condiciones y bajo qué marcos normativos y técnicos se producirá esa incorporación. La evidencia disponible sugiere que la diferencia entre una incorporación que profundice las desigualdades y una que contribuya a reducirlas no depende principalmente de la tecnología en sí misma, sino de las decisiones institucionales, regulatorias y políticas que rodean su adopción.

Schwartz et al. (2020) acuñaron la distinción entre "Red AI" IA que prioriza el rendimiento técnico a cualquier costo computacional y social y "Green AI" IA que busca la eficiencia técnica con consideración explícita de sus costos ambientales y sociales. Esta dicotomía conceptual puede extenderse útilmente al contexto colombiano: la elección entre una adopción de IA que ponga al servicio de los grupos más poderosos las capacidades de estos sistemas y una adopción que busque activamente distribuir sus beneficios y minimizar sus daños para los grupos más vulnerables es fundamentalmente una elección política, no técnica.

La ausencia de Colombia del debate global sobre IA ética y sostenible es, en última instancia, una forma de delegación epistémica y política: al no participar activamente en la producción del conocimiento y de los marcos normativos que gobiernan el desarrollo de la IA, Colombia delega en otros actores las decisiones sobre qué valores se codifican en las tecnologías que ella adopta, qué poblaciones son estudiadas y protegidas, y qué definiciones de equidad y sostenibilidad prevalecen. Esta delegación tiene costos que,

como sugiere la evidencia del corpus, recaen de manera desproporcionada sobre los grupos más vulnerables de la sociedad colombiana.

Revertirla requiere no solo inversión en infraestructura técnica y formación de talento, sino, fundamentalmente, la construcción de una perspectiva epistémica propia sobre la IA crítica, contextualizada y comprometida con la justicia que sitúe a Colombia como sujeto activo, y no solo objeto pasivo, del orden tecnológico contemporáneo.

5 CONCLUSIONES

La presente revisión sistemática de literatura, realizada sobre un corpus de 130 fuentes de alto impacto publicadas entre 2016 y 2026, permite articular un conjunto de conclusiones que trascienden la suma de sus partes: una lectura integrada del estado actual del desarrollo de la Inteligencia Artificial Generativa (IAG) desde las dimensiones ética, técnica, sectorial y ambiental. Estas conclusiones se organizan en torno a las cinco preguntas que articularon el diseño investigativo, y se proyectan hacia las implicaciones para el contexto colombiano y latinoamericano.

5.1 Los sesgos de la IAG son estructurales, no accidentales

El hallazgo más consistente y transversalmente respaldado por el corpus es que los sesgos identificados en los sistemas de Inteligencia Artificial Generativa no constituyen errores técnicos aislados ni defectos corregibles mediante ajustes marginales en los algoritmos. Son, en cambio, expresiones estructurales de las asimetrías de poder, representación y acceso que caracterizan las sociedades donde estos sistemas son desarrollados. Esta conclusión, articulada desde perspectivas disciplinarias tan diversas como la filosofía (Henao Henao y Hernández Mora, 2025), la informática médica (Chen et al., 2024), el aprendizaje automático (Pessach y Shmueli, 2022) y el análisis jurídico (Casanova, 2025), representa el consenso más sólido del campo.

Los seis tipos de sesgo identificados en la literatura de representación, de medición, de evaluación, de implementación, específicos de LLMs e histórico-estructurales no operan de manera independiente sino en interacción dinámica. Un sesgo de representación en los datos de entrenamiento genera un sesgo de medición en las métricas de rendimiento, que a su vez produce un sesgo de evaluación cuando los conjuntos de prueba tampoco son representativos, y todo ello se amplifica en la fase de implementación cuando los

operadores humanos tienden a validar las recomendaciones del sistema que confirman sus prejuicios previos (Selten et al., 2023). Esta circularidad sistémica implica que la identificación y mitigación de los sesgos requiere intervenciones simultáneas en múltiples niveles del ciclo de vida del modelo, no soluciones puntuales en una sola etapa.

En los sectores de salud y justicia donde las consecuencias de los sesgos son más directas e irreversibles la evidencia es especialmente contundente. En salud, modelos entrenados con registros electrónicos tienden a confundir el acceso diferencial a la atención con la necesidad médica real, perjudicando precisamente a los grupos más vulnerables (Chen et al., 2024; De Andrade et al., 2026). En justicia, sistemas como COMPAS reproducen disparidades raciales históricas bajo una apariencia de objetividad técnica, asignando sistemáticamente mayores puntajes de riesgo a personas de grupos étnicos históricamente criminalizados (Pessach y Shmueli, 2022; Alikhademi et al., 2022). La conclusión es inequívoca: no es posible construir sistemas de IA equitativos sobre datos producidos en condiciones de desigualdad estructural sin una teoría normativa explícita de lo que constituye una sociedad justa y sin mecanismos activos de corrección.

5.2 La calidad de los datos de entrenamiento es condición necesaria, pero no suficiente, para la equidad

La segunda conclusión mayor de esta revisión concierne al papel determinante y frecuentemente subestimado de la calidad de los datos de entrenamiento en el comportamiento ético de los sistemas de IAG. El principio de garbage in, garbage out adquiere en el aprendizaje profundo una dimensión adicional: los modelos no solo reproducen las deficiencias de sus datos, sino que las amplifican de manera no lineal y a escala masiva.

De Andrade et al. (2026) documentan que en bases de datos de cuidados críticos algunas de las más utilizadas en investigación de IA médica las tasas de datos faltantes superan el 80% para algunas variables críticas, los errores de medicación afectan hasta el 1,6% de los registros, y el 40% de las características predictivas son indicadores de ausencia de datos en lugar de valores clínicos reales. Vajpayee y Khobragade (2024) señalan que los ensayos clínicos fuente primaria de datos de alta calidad han excluido históricamente a mujeres, personas mayores y minorías étnicas, comprometiendo la base de evidencia sobre la que se construyen los modelos de IA en especialidades críticas. Chen

et al. (2024) advierten que la interoperabilidad técnica de los sistemas de información no garantiza automáticamente la representatividad equitativa: se requieren esfuerzos adicionales y explícitos para asegurar que los datos provengan de poblaciones diversas.

La conclusión clave en esta dimensión es que la calidad de los datos es condición necesaria pero no suficiente para la equidad. Incluso con datos técnicamente correctos precisos, completos y consistentes, si su distribución refleja desigualdades estructurales preexistentes, los modelos construidos sobre ellos heredarán y amplificarán esas desigualdades. La mejora de la calidad de los datos requiere, por tanto, no solo gobernanza técnica de los sistemas de información, sino también políticas activas de inclusión representativa que corrijan los desequilibrios históricos en la producción y registro de datos sobre distintos grupos poblacionales.

5.3 La efectividad de las estrategias de mitigación es real, pero limitada y contextuales

La tercera conclusión de la revisión atañe al estado de las estrategias de mitigación de sesgos disponibles. El corpus revela un campo con notable proliferación de propuestas técnicas de preprocesamiento, algoritmos con restricciones de equidad, marcos de postprocesamiento, arquitecturas de gobernanza y con una evidencia de efectividad que es sistemáticamente moderada, contextual y acompañada de trade-offs no resueltos.

Skaiky et al. (2025) demuestran empíricamente que mejorar la equidad intergrupal en un sistema de predicción de riesgo implica una reducción de 10 puntos porcentuales en la exactitud global. Chand et al. (2026) prueban que reducir un tipo de sesgo en LLMs tiende a amplificar otros: intervenir sobre el sesgo de género puede intensificar el sesgo racial y viceversa. Pessach y Shmueli (2022) documentan la imposibilidad matemática de satisfacer simultáneamente varias definiciones formales de equidad cuando los grupos difieren en tasas base de la variable de resultado. Este conjunto de hallazgos configura lo que puede denominarse el dilema central de la mitigación: no existe una solución técnica universal que resuelva todos los tipos de sesgo simultáneamente sin costo en rendimiento o en otras dimensiones de la equidad.

La implicación de esta conclusión para la práctica es significativa: la elección entre distintas estrategias de mitigación no es solo técnica sino normativa. Decidir qué tipo de equidad priorizar paridad estadística, igualdad de oportunidad, calibración es una decisión

que implica valores y que tiene ganadores y perdedores. Esta decisión no puede ser delegada exclusivamente a los equipos técnicos que desarrollan los modelos: requiere deliberación pública, participación de las comunidades afectadas y rendición de cuentas institucional. Los marcos de gobernanza como NoBIAS (Alvarez et al., 2024) o los propuestos por Papagiannidis et al. (2025) son prometedores precisamente porque articulan la dimensión técnica con la organizacional y la regulatoria, aunque la evidencia de su efectividad en implementaciones reales sigue siendo escasa.

5.4 La sostenibilidad ambiental de la IAG es una urgencia sistémica infraatendida

La cuarta conclusión mayor de la investigación concierne a la dimensión ambiental del desarrollo de la IAG, que el corpus identifica sistemáticamente como el eje más subestimado tanto en la investigación como en la práctica. Los datos cuantitativos disponibles dibujan un panorama de consumo energético extraordinario: el entrenamiento de GPT-3 consumió 1.287 MWh y generó 552 toneladas de CO_2 equivalente (Alzoubi y Mishra, 2024); ChatGPT consume aproximadamente 260 MWh por día solo en inferencia (Bolón-Canedo et al., 2024); y los costos computacionales para entrenar modelos de vanguardia se multiplicaron por 300.000 entre 2012 y 2017 (Schwartz et al., 2020).

Lo más preocupante no es la magnitud absoluta de estos números, sino la dinámica que los genera: los incentivos académicos e industriales actuales favorecen el escalamiento de modelos en la búsqueda de mejoras marginales de rendimiento, sin consideración proporcional de los costos ambientales. Schwartz et al. (2020) documentaron que ganar un punto porcentual de mejora en algunos benchmarks puede requerir multiplicar por cien el costo computacional, lo que apunta a rendimientos decrecientes que hacen la trayectoria de escalamiento insostenible en términos ambientales.

El paradigma de la IA Verde (Verdecchia et al., 2023) ofrece un repertorio de estrategias técnicas poda de modelos, cuantización, destilación del conocimiento, hardware especializado que pueden reducir el consumo energético entre un 13% y un 115% dependiendo de la técnica y el contexto. Bolón-Canedo et al. (2024) y Dorrich et al. (2023) demuestran que técnicas de precisión mixta son accesibles con las herramientas disponibles en los principales frameworks de aprendizaje profundo. Sin embargo, la evidencia disponible también señala que solo el 23% de los estudios sobre IA Verde

involucran socios industriales (Verdecchia et al., 2023), revelando una brecha significativa entre el conocimiento técnico disponible y su adopción en la práctica.

La conclusión en esta dimensión es doble: la sostenibilidad ambiental no es un criterio de segundo orden añadible a posteriori al desarrollo de los modelos, sino una condición de viabilidad del paradigma de la IAG a largo plazo; y la paradoja identificada por Gaur et al. (2023) que la IA puede ser tanto una de las herramientas más poderosas para abordar la crisis climática como una de sus fuentes más dinámicas solo se resolverá si el campo adopta la eficiencia ambiental como métrica de éxito al mismo nivel que el rendimiento técnico.

5.5 La gobernanza efectiva requiere marcos vinculantes, capacidades institucionales y participación ciudadana

La quinta conclusión transversal de esta revisión concierne a las condiciones de la gobernanza efectiva de la IAG. El corpus analizado converge en que los marcos de principios voluntarios por más sofisticados que sean conceptualmente son insuficientes para garantizar prácticas de desarrollo responsable. Incluso el EU AI Act, el marco regulatorio más avanzado disponible, enfrenta desafíos significativos de implementación relacionados con la capacidad técnica de las entidades supervisoras, la complejidad de la evaluación de sistemas de IA de alto riesgo y la velocidad del desarrollo tecnológico.

González-Fernández et al. (2025) evidencian que la brecha entre los marcos normativos adoptados y las prácticas institucionales reales continúa siendo considerable, en parte por la falta de herramientas operativas y en parte por la insuficiente formación de los actores responsables de la implementación. Papagiannidis et al. (2025) documentan que las organizaciones que implementan prácticas estructurales, procedimentales y relacionales de manera integrada presentan mejores resultados de equidad, pero que este enfoque comprehensivo es poco frecuente en la práctica. Decker et al. (2025) aportan evidencia de que la participación ciudadana en el diseño y evaluación de sistemas algorítmicos no solo mejora su legitimidad democrática sino también su equidad técnica.

Tres condiciones emergen del corpus como necesarias aunque no suficientes para una gobernanza efectiva de la IAG: primero, marcos regulatorios vinculantes que generen obligaciones y mecanismos de supervisión, no solo recomendaciones voluntarias; segundo, capacidades institucionales técnicas que permitan a los supervisores evaluar realmente los

sistemas que regulan, en lugar de depender de la autorregulación de los desarrolladores; y tercero, mecanismos formales de participación de las comunidades afectadas en todas las etapas del ciclo de vida de los sistemas, desde la definición de los datos de entrenamiento hasta la evaluación de los resultados en producción. La ausencia de cualquiera de estas tres condiciones genera una gobernanza que es, en el mejor de los casos, simbólica.

5.6 Implicaciones para el contexto colombiano y latinoamericano: una agenda urgente

Las conclusiones anteriores adquieren una dimensión específica cuando se leen desde el contexto colombiano y latinoamericano. La región no está al margen del desarrollo de la IAG: los sistemas desarrollados fundamentalmente en el norte global se despliegan de manera creciente en contextos colombianos en salud, en educación, en servicios financieros, en sistemas de seguridad llevando consigo los sesgos y las limitaciones documentadas en el corpus, pero sin los marcos regulatorios, las capacidades de evaluación ni los mecanismos de participación ciudadana que podrían mitigar sus efectos negativos.

Como se ha expuesto, la dependencia de modelos de IA desarrollados con datos ajenos al contexto nacional exacerba el riesgo de fallo de transporte algorítmico en áreas sensibles (Calahorrano Latorre, 2025; Goetz et al., 2024). La falta de sintonía entre los sistemas globales y la realidad colombiana visible tanto en la desatención de perfiles de salud locales como en la inoperancia de los modelos frente al patrimonio lingüístico y las lenguas indígenas subraya una vulnerabilidad estructural. No se trata solo de una limitación técnica, sino de una barrera que compromete la equidad en el acceso a los beneficios de la inteligencia artificial.

La ausencia de Colombia del debate global sobre IA ética y sostenible documentada por el hecho de que el 70% de los investigadores sobre sesgos en LLMs trabajan en instituciones de los Estados Unidos (Ducel et al., 2023) no es una omisión neutral. Es una forma de delegación epistémica y política que tiene consecuencias distributivas concretas: los valores codificados en las tecnologías adoptadas, las definiciones de equidad que prevalecen, las poblaciones que son estudiadas y protegidas, son determinadas por actores externos con intereses y perspectivas que no necesariamente coinciden con los de la sociedad colombiana.

La agenda que emerge de esta investigación para el contexto colombiano puede articularse en cuatro prioridades interrelacionadas: (1) el desarrollo de investigación de validación cruzada de sistemas de IA en uso en Colombia, para documentar empíricamente las disparidades de rendimiento entre grupos demográficos que el corpus sugiere son altamente probables; (2) la construcción de marcos normativos vinculantes adaptados al contexto jurídico e institucional colombiano, tomando como referencia el EU AI Act pero con criterios de proporcionalidad realistas para la capacidad supervisora disponible; (3) la articulación de la dimensión ambiental en la política de IA colombiana, considerando las implicaciones del consumo energético del cómputo para los compromisos climáticos del país; y (4) el fortalecimiento de capacidades de investigación e innovación en IA que permitan a Colombia participar como productor y no solo como consumidor de las tecnologías y los marcos normativos que gobiernan el campo.

5.7 Reflexión final: el imperativo de una IA éticamente comprometida

Esta revisión sistemática comenzó con una pregunta sobre cómo influyen los sesgos cognitivos, la calidad de los datos y el impacto ambiental en el desarrollo ético y sostenible de la IAG. La respuesta que emerge del corpus es, en síntesis, la siguiente: todos estos factores influyen de manera determinante, sus efectos son interactivos y no aditivos, y ninguno puede ser abordado de manera efectiva sin los otros.

La IAG ética y sostenible no es un estado que se alcanza mediante la eliminación de todos los sesgos, la garantía de calidad perfecta en los datos o la reducción a cero de la huella de carbono. Es un proceso continuo de mejora, auditoría, rendición de cuentas y adaptación, orientado por compromisos normativos explícitos y respaldado por instituciones con capacidad real de supervisión. La IA Confiable que es legal, ética y robusta (Li et al., 2023) no es el punto de partida del desarrollo tecnológico, sino su objetivo permanente.

El desafío más profundo que revela esta investigación no es técnico sino político: requiere que las sociedades y en particular las sociedades latinoamericanas que hasta ahora han participado marginalmente en este debate tomen decisiones colectivas sobre qué valores deben guiar el desarrollo de tecnologías que afectan profundamente la vida de todos sus miembros. La investigación académica rigurosa, como la que esta revisión sistemática pretende aportar, tiene el rol de producir la evidencia y los marcos conceptuales que hagan esas decisiones más informadas, más conscientes de sus implicaciones y más

abiertas a la deliberación democrática. Es, en última instancia, una contribución a la construcción de una sociedad donde la tecnología sirve a la justicia, y no al revés.

REFERENCIAS

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N. P., Parwani, A. v., Gibbs, J., Agosto-Arroyo, E., Beck, A. H., & Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *Journal of Pathology*, *249*(3), 286–294. https://doi.org/10.1002/path.5331

Abràmoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., Eydelman, M. B., & Maisel, W. H. (2023). Considerations for addressing bias in artificial intelligence for health equity. *Npj Digital Medicine*, *6*(1). https://doi.org/10.1038/s41746-023-00913-9

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, *12*(1). https://doi.org/10.1016/j.hlpt.2022.100702

Ahadian, P., Xu, W., Liu, D., & Guan, Q. (2026). Ethics of trustworthy AI in healthcare: Challenges, principles, and practical pathways. *Neurocomputing*, *661*, 131942. https://doi.org/10.1016/j.neucom.2025.131942

Ahmad, A., Vallès, Y., & Idaghdour, Y. (2026). Bias in AI systems: integrating formal and socio-technical approaches. *Frontiers in Big Data*, *8*. https://doi.org/10.3389/fdata.2025.1686452

Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J. E. (2022). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, *30*(1), 1–17. https://doi.org/10.1007/s10506-021-09286-4

Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology*, *26*(2). https://doi.org/10.1007/s10676-024-09746-w

Alzoubi, Y. I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. *Journal of Cleaner Production*, *468*. https://doi.org/10.1016/j.jclepro.2024.143090

Arias-Montaño, E. I., León-Japa, R. S., García-Jaramillo, P., & Maldonado-Ortega, J. (2026). ESTIMACIÓN DE EMISIONES POR FALLOS EN MOTORES OTTO MEDIANTE REDES NEURONALES CONVOLUCIONALES. *Ingenius, Revista Ciencia y Tecnología*, (35), 97–109. https://doi.org/10.17163/ings.n35.2026.07

Asaro, P. M. (2019). AI ethics in predictive policing: From models of threat to an ethics of care. *IEEE Technology and Society Magazine*, *38*(2), 40–53. https://doi.org/10.1109/MTS.2019.2915154

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, *122*(8). https://doi.org/10.1073/pnas.2416228122

Berk, R. A. (2021). Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement. *Annual Review of Criminology*, *4*(1), 209–237. https://doi.org/10.1146/annurev-criminol-051520-012342

Bissoto, A., Barata, C., Valle, E., & Avila, S. (2024). Even small correlation and diversity shifts pose dataset-bias issues. *Pattern Recognition Letters*, *179*, 87–93. https://doi.org/10.1016/J.PATREC.2024.01.026

Bolón-Canedo, V., Morán-Fernández, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. *Neurocomputing*, *599*. https://doi.org/10.1016/j.neucom.2024.128096

Bouza, L., Bugeau, A., & Lannelongue, L. (2023). How to estimate carbon footprint when training deep learning models? A guide and review. *Environmental Research Communications*, *11*(115). https://doi.org/10.1088/2515-7620/ACF81B

Brayne, S. (2020). Predict and surveil: Data, discretion, and the future of policing. In *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press. https://doi.org/10.1093/oso/9780190684099.001.0001

Brown, K. E., Yan, C., Li, Z., Zhang, X., Collins, B. X., Chen, Y., Clayton, E. W., Kantarcioglu, M., Vorobeychik, Y., & Malin, B. A. (2025). Large language models are less effective at clinical prediction tasks than locally trained machine learning models. *Journal of*

the American Medical Informatics Association, *32*(5), 811–822. https://doi.org/10.1093/jamia/ocaf038

Budennyy, S. A., Lazarev, V. D., Zakharenko, N. N., Korovin, A. N., Plosskaya, O. A., Dimitrov, D. v., Akhripkin, V. S., Pavlov, I. v., Oseledets, I. v., Barsola, I. S., Egorov, I. v., Kosterina, A. A., & Zhukov, L. E. (2022). eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. *Doklady Mathematics*, *106*, S118–S128. https://doi.org/10.1134/S1064562422060230

Calahorrano Latorre, E. (2025). Equidad algorítmica y decisiones clínicas basadas en sistemas de soporte basados en inteligencia artificial. *Nuevo Derecho*, *21*(36), 1–17. https://doi.org/10.25057/2500672X.1691

Cao, Y., Yang, H., Xue, Y., Wang, F., Li, T., Zhao, L., & Fu, Y. (2026). Ethical risks and considerations of the integration of Brain-Computer Interfaces with Artificial Intelligence. *Cognitive Neurodynamics*, *20*(1). https://doi.org/10.1007/s11571-025-10380-5

CAPELLÀ I. RICART, A. (2026). TRATAMIENTO DE CATEGORIAS ESPECIALES DE DATOS PARA LA DETECCIÓN Y CORRECCIÓN DE SESGOS EN SISTEMAS DE IA DE ALTO RIESGO: ¿UNA MEDIDA DE ACCIÓN POSITIVA? *Revista Derechos y Libertades*, (54), 113–137. https://doi.org/10.20318/dyl.2026.9974

CASANOVA, E. A. (2025). BIAS AND DISCRIMINATION IN ARTIFICIAL INTELLIGENCE SYSTEMS, CAUSES AND LEGAL RESPONSES IN THE EUROPEAN UNION. *IgualdadES*, (13), 181–208. https://doi.org/10.18042/cepc/igdes.13.07

Castellanos-Nieves, D., & García-Forte, L. (2023). Improving Automated Machine-Learning Systems through Green AI. *Applied Sciences (Switzerland)*, *13*(20). https://doi.org/10.3390/APP132011583

Castellanos-Nieves, D., & García-Forte, L. (2024). Strategies of Automated Machine Learning for Energy Sustainability in Green Artificial Intelligence. *Applied Sciences (Switzerland)*, *14*(14). https://doi.org/10.3390/APP14146196

Chand, S., Baca, F., & Ferrara, E. (2026). No Free Lunch in Language Model Bias Mitigation? Targeted Bias Reduction Can Exacerbate Unmitigated LLM Biases. *AI (Switzerland)*, *7*(1), 24. https://doi.org/10.3390/ai7010024

Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic

health record-based models. *Journal of the American Medical Informatics Association*, *31*(5), 1172–1183. https://doi.org/10.1093/jamia/ocae060

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

Chen, X., Zhang, J., Lin, B., Chen, Z., Wolter, K., & Min, G. (2022). Energy-Efficient Offloading for DNN-Based Smart IoT Systems in Cloud-Edge Environments. *IEEE Transactions on Parallel and Distributed Systems*, *33*(3), 683–697. https://doi.org/10.1109/TPDS.2021.3100298

Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., Jain, A., Jordan, W. B., Konya, S., Moore, R. H., Moore, T. T., Rodriguez, R., Shaheen, G., Snyder, L. P., Srinivasan, M., … Ohno-Machado, L. (2023). Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Network Open*, *6*(12), E2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050

Choudhary, T. (2025). Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude. *IEEE Access*, *13*, 11341–11379. https://doi.org/10.1109/ACCESS.2024.3523764

Cobert, J., Mills, H., Lee, A., Gologorskaya, O., Espejo, E., Jeon, S. Y., Boscardin, W. J., Heintz, T. A., Kennedy, C. J., Ashana, D. C., Chapman, A. C., Raghunathan, K., Smith, A. K., & Lee, S. J. (2024). Measuring Implicit Bias in ICU Notes Using Word-Embedding Neural Network Models. *CHEST*, *165*(6), 1481–1490. https://doi.org/10.1016/J.CHEST.2023.12.031

de Andrade, J. B. C., de Medeiros Cavalcante, M. A. O., Lopes, T. L. M., Treigher, J. M. S., Balsells, M. D., Vasconcelos, J. L., Monteiro, L. C., & da Silveira Mota, D. D. (2026). Discovery of data quality issues in electronic health records: profound consequences for critical care medicine applications – a systematized review. *Critical Care*, *30*(1). https://doi.org/10.1186/s13054-025-05677-0

Decker, M. C., Wegner, L., & Leicht-Scholten, C. (2025). Procedural fairness in algorithmic decision-making: the role of public engagement. *Ethics and Information Technology*, *27*(1). https://doi.org/10.1007/s10676-024-09811-4

Díaz-Rodríguez, N., del Ser, J., Coeckelbergh, M., López de Prado, M., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Information Fusion*, *99*. https://doi.org/10.1016/j.inffus.2023.101896

Dorrich, M., Fan, M., & Kist, A. M. (2023). Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks. *IEEE Access*, *11*, 57627–57634. https://doi.org/10.1109/ACCESS.2023.3284388

Ducel, F., Neveol, A., & Fort, K. (2023, January 1). *Bias Research for Language Models is Biased: a Survey for Deconstructing Bias in Large Language Models-Web of Science Core Collection*. TRAITEMENT AUTOMATIQUE DES LANGUES. https://doi.org/WOS:001494962200006

Fiske, A., Radhuber, I. M., Willem, T., Buyx, A., Celi, L. A., & McLennan, S. (2025). Climate change and health: the next challenge of ethical AI. *The Lancet Global Health*, *13*(7), e1314–e1320. https://doi.org/10.1016/S2214-109X(25)00124-X

Gallese, C. (2026). Predictive policing and predictive justice: Ethics, data protection, and the AI act. *Computer Law & Security Review*, *61*, 106282. https://doi.org/10.1016/J.CLSR.2026.106282

Gameiro, R. R., Woite, N. L., Sauer, C. M., Hao, S., Fernandes, C. O., Premo, A. E., Teixeira, A. R., Resli, I., Wong, A. K. I., & Celi, L. A. (2025). The Data Artifacts Glossary: a community-based repository for bias on health datasets. *Journal of Biomedical Science*, *32*(1), 1–9. https://doi.org/10.1186/S12929-024-01106-6

Gaur, L., & Abraham, A. (2026). Understanding training data and mitigating biases in training data. *Generative Artificial Intelligence and Ethics for Healthcare*, 25–38. https://doi.org/10.1016/B978-0-443-33124-4.00008-4

Gaur, L., Afaq, A., Arora, G. K., & Khan, N. (2023). Artificial intelligence for carbon emissions using system of systems theory. *Ecological Informatics*, *76*, 102165. https://doi.org/10.1016/J.ECOINF.2023.102165

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI IN IMAGING AND THERAPY: INNOVATIONS, ETHICS, AND IMPACT: REVIEW ARTICLE AI pitfalls and what not to do: mitigating bias in AI. *British Journal of Radiology*, *96*(1150). https://doi.org/10.1259/bjr.20230023

Glocker, B., Jones, C., Bernhardt, M., & Winzeck, S. (2023). Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *eBioMedicine*, *89*. https://doi.org/10.1016/j.ebiom.2023.104467

Goetz, L., Seedat, N., Vandersluis, R., & van der Schaar, M. (2024). Generalization—a key challenge for responsible AI in patient-facing clinical applications. *Npj Digital Medicine*, *7*(1). https://doi.org/10.1038/s41746-024-01127-3

Goktas, P., & Grzybowski, A. (2025). Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI. *Journal of Clinical Medicine*, *14*(5). https://doi.org/10.3390/jcm14051605

González-Fernández, M. O., Romero-López, M. A., Sgreccia, N. F., & Latorre Medina, M. J. (2025). Marcos normativos para una IA ética y confiable en la educación superior: estado de la cuestión. *RIED: Revista Iberoamericana de Educación a Distancia*, *28*(2), 181–208. https://doi.org/10.5944/ried.28.2.43511

Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *Minds and Machines*, *34*(4), 39. https://doi.org/10.1007/s11023-024-09694-w

Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *Npj Digital Medicine*, *8*(1). https://doi.org/10.1038/s41746-025-01503-7

Haxvig, H. A., D'Andrea, V., & Teli, M. (2025). "I've never seen a glass ceiling better represented": Bias and gendering in LLM-generated synthetic personas from a participatory design perspective. *International Journal of Human Computer Studies*, *205*. https://doi.org/10.1016/j.ijhcs.2025.103651

Henao Henao, J., & Hernández Mora, L. H. (2025). LA INTELIGENCIA ARTIFICIAL Y SUS IMPLICACIONES ÉTICAS EN LA ERA DE LOS GRANDES MODELOS DE LENGUAJE (LLM). *Praxis Filosófica*, (62S), 1–25. https://doi.org/10.25100/pfilosofica.v0i62s.15442

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, *2018*(8). https://doi.org/10.7717/peerj.5518

Hills, T. T. (2026). Could You Be Wrong: Metacognitive Prompts for Improving Human Decision Making Help LLMs Identify Their Own Biases. *AI (Switzerland)*, *7*(1), 33. https://doi.org/10.3390/ai7010033

Hossain, M., Khalid, H., Rao, A. P., Lootah, M., Al-Mohammedi, S. S. K., & Majeed, S. R. (2024). Comprehensive Review of AI, IoT, and ML in Enhancing Urban Mobility and Reducing Carbon Footprints. *2024 3rd International Conference on Sustainable Mobility Applications, Renewables and Technology, SMART 2024*. https://doi.org/10.1109/SMART63170.2024.10815521

Hu, Y., Dai, L., Wu, Z., Huang, W., Liu, J., Li, Y., & Yan, Y. (2025). Publishing Ethics and AI: Challenges and Opportunities for Algorithm-Generated Content. *LatIA*, *3*, 1–13. https://doi.org/10.62486/LATIA2025378

Inoshita, K. (2024). The Efficient Development of Conflict Structure Datasets for Evaluating Sentiment Recognition Bias in Large Language Models. *Proceedings of the International Conference on Electrical Engineering and Informatics*, 7–12. https://doi.org/10.1109/ICELTICS62730.2024.10776050

Jaradat, S., Acharya, N., Shivshankar, S., Alhadidi, T. I., & Elhenawy, M. (2025). AI for Data Quality Auditing: Detecting Mislabeled Work Zone Crashes Using Large Language Models. *Algorithms*, *18*(6), 317. https://doi.org/10.3390/A18060317

Karthikeyan, R., & Boudourides, M. (2026). The algorithmic blind spot: bias, moral status, and the future of robot rights. *AI & Society*, 1–11. https://doi.org/10.1007/S00146-026-03003-Y

Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive policing and the politics of patterns. *British Journal of Criminology*, *59*(3), 674–692. https://doi.org/10.1093/bjc/azy060

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, *55*(2). https://doi.org/10.1145/3491209

Kounadi, O., Ristea, A., Araujo, A., & Leitner, M. (2020). A systematic review on spatial crime forecasting. *Crime Science*, *9*(1). https://doi.org/10.1186/s40163-020-00116-7

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3555803

Li, S., & Deng, W. (2022). A Deeper Look at Facial Expression Dataset Bias. *IEEE Transactions on Affective Computing*, *13*(2), 881–893. https://doi.org/10.1109/TAFFC.2020.2973158

Li, X., Ye, P., Li, J., Liu, Z., Cao, L., & Wang, F. Y. (2022). From Features Engineering to Scenarios Engineering for Trustworthy AI: I&I, C&C, and V&V. *IEEE Intelligent Systems*, *37*(4), 18–26. https://doi.org/10.1109/MIS.2022.3197950

Linera, M. Á. P. (2023). Policía predictiva y prevención de la violencia de género: el sistema VioGén. *Revista de Internet, Derecho y Política (IDP)*, (39), 1–13. https://doi.org/10.7238/IDP.V0I39.416473

Liu, M., Ning, Y., Teixayavong, S., Liu, X., Mertens, M., Shang, Y., Li, X., Miao, D., Liao, J., Xu, J., Ting, D. S. W., Cheng, L. T.-E., Ong, J. C. L., Teo, Z. L., Tan, T. F., RaviChandran, N., Wang, F., Celi, L. A., Ong, M. E. H., & Liu, N. (2025). A scoping review and evidence gap analysis of clinical AI fairness. *Npj Digital Medicine*, *8*(1), 360. https://doi.org/10.1038/s41746-025-01667-2

Lou, J., & Sun, Y. (2026). Anchoring bias in large language models: an experimental study. *Journal of Computational Social Science*, *9*(1). https://doi.org/10.1007/s42001-025-00435-2

Luo, L., Huang, X., Wang, M., Wan, Z., Ma, W., & Chen, H. (2025). Mitigating medical dataset bias by learning adaptive agreement from a biased council. *Medical Image Analysis*, *105*, 103629. https://doi.org/10.1016/J.MEDIA.2025.103629

Lyu, M., Yang, Z., Chen, C., & Huang, J. (2026). Does AI prefer healthy foods? Examining value bias in large language models. *International Journal of Gastronomy and Food Science*, *43*. https://doi.org/10.1016/j.ijgfs.2025.101406

mac Namee, B., Cunningham, P., Byrne, S., & Corrigan, O. I. (2002). The problem of bias in training data in regression problems in medical decision support. *Artificial Intelligence in Medicine*, *24*(1), 51–70. https://doi.org/10.1016/S0933-3657(01)00092-6

Mana, A. A., Allouhi, A., Hamrani, A., Rahman, S., el Jamaoui, I., & Jayachandran, K. (2024). Sustainable AI-based production agriculture: Exploring AI applications and implications in agricultural practices. *Smart Agricultural Technology*, *7*. https://doi.org/10.1016/j.atech.2024.100416

Martínez-Rolán, X., Cabezuelo-Lorenzo, F., & Oliveira, L. (2025). The new scenarios of the digital society in the face of the challenge of Generative Artificial Intelligence (AI)[Os

novos cenários da sociedade digital diante do desafio da Inteligência Artificial (IA) Gerativa][LOS NUEVOS ESCENARIOS DE LA SOCIEDAD DIGITAL ANTE EL DESAFÍO DE LA INTELIGENCIA ARTIFICIAL (IA) GENERATIVA]. *Encontros Bibli*, *30*. https://doi.org/10.5007/1518-2924.2025.e105080

Mataraso, S., D'Souza, S., Seong, D., Berson, E., Espinosa, C., & Aghaeepour, N. (2025). Benchmarking of pre-training strategies for electronic health record foundation models. *JAMIA Open*, *8*(4). https://doi.org/10.1093/JAMIAOPEN/OOAF090

Meijer, A., Lorenz, L., & Wessels, M. (2021). Algorithmization of Bureaucratic Organizations: Using a Practice Lens to Study How Context Shapes Predictive Policing Systems. *Public Administration Review*, *81*(5), 837–846. https://doi.org/10.1111/puar.13391

Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration*, *42*(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

Melak, A., Aseged, T., & Shitaw, T. (2024). The Influence of Artificial Intelligence Technology on the Management of Livestock Farms. *International Journal of Distributed Sensor Networks*, *2024*. https://doi.org/10.1155/2024/8929748

MELO, Z. (2024). Inteligencia artificial en salud: desafíos éticos para lograr la aplicación de las tecnologías a la salud del paciente. *TraHs - Trayectorias Humanas Trascontinentales*, (18), 49–62. https://doi.org/10.25965/trahs.6349

Mhlambi, S., & Tiribelli, S. (2023). Decolonizing AI Ethics: Relational Autonomy as a Means to Counter AI Harms. *Topoi*, *42*(3), 867–880. https://doi.org/10.1007/s11245-022-09874-2

Mir, B. A., Abbas, S. R., & Lee, S. W. (2026). Federated Learning in Healthcare Ethics: A Systematic Review of Privacy-Preserving and Equitable Medical AI. *Healthcare (Basel, Switzerland)*, *14*(3). https://doi.org/10.3390/HEALTHCARE14030306

Moghimi, S. M., Gulliver, T. A., & Thirumai Chelvan, I. (2024). Energy Management in Modern Buildings Based on Demand Prediction and Machine Learning—A Review. *Energies*, *17*(3). https://doi.org/10.3390/en17030555

Mouri Zadeh Khaki, A., & Choi, A. (2026). Evaluating Fairness in LLM Negotiator Agents via Economic Games Using Multi-Agent Systems. *Mathematics (2227-7390)*, *14*(3), 458. https://doi.org/10.3390/MATH14030458

Muhammad Kakale, M., & Pinelli, N. (2026). Impacts of artificial intelligence on healthcare business models and outcomes. *Journal of Health Organization and Management*, *40*(9), 39–74. https://doi.org/10.1108/JHOM-06-2025-0356

Muñoz-García, V., Consuegra-Ayala, J. P., & Moreda, P. (2025). Leading and non-leading prompts: Quantifying gender bias in Large Language Models through BiasBloom corpus. *Knowledge-Based Systems*, *325*. https://doi.org/10.1016/j.knosys.2025.113915

Nguyen, K. V. (2025). The Use of Generative AI Tools in Higher Education: Ethical and Pedagogical Principles. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-025-09607-1

Nomelini, G. G., & Marcolin, C. B. (2024). Gender bias in large language models: A job postings analysis. *RAM. Mackenzie Management Review / RAM. Revista de Administração Mackenzie*, *25*(6), 1–27. https://doi.org/10.1590/1678-6971/eRAMD240056

Nouis, S. C. E., Uren, V., & Jariwala, S. (2025). Evaluating accountability, transparency, and bias in AI-assisted healthcare decision- making: a qualitative study of healthcare professionals' perspectives in the UK. *BMC Medical Ethics*, *26*(1). https://doi.org/10.1186/s12910-025-01243-z

Nyawambi, T. M., & Muchiri, H. (2024). Mitigating Racial Algorithmic Bias in Healthcare Artificial Intelligent Systems: A Fairness-Aware Machine Learning Approach. *2024 5th International Conference on Smart Sensors and Application: Shaping the Future of Intelligent Innovation, ICSSA 2024*. https://doi.org/10.1109/ICSSA62312.2024.10788666

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *Journal of Strategic Information Systems*, *34*(2). https://doi.org/10.1016/j.jsis.2024.101885

Parra, C. M., Gupta, M., & Dennehy, D. (2022). Likelihood of Questioning AI-Based Recommendations Due to Perceived Racial/Gender Bias. *IEEE Transactions on Technology and Society*, *3*(1), 41–45. https://doi.org/10.1109/TTS.2021.3120303

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3). https://doi.org/10.1145/3494672

Polevikov, S. (2023). Advancing AI in healthcare: A comprehensive review of best practices. *Clinica Chimica Acta*, *548*. https://doi.org/10.1016/j.cca.2023.117519

Pryor, K., Coleman, T., Hossain, S. Z., Tootil, E., & Ahmed, S. (2025). Examining Large Language Models Within Autism-Related Contexts: A Systematic Review of Bias and

(Mis) Representation. *Proceedings - 2025 IEEE 49th Annual Computers, Software, and Applications Conference, COMPSAC 2025*, 876–886. https://doi.org/10.1109/COMPSAC65507.2025.00115

Rajawat, A. S., Goyal, S. B., & Abdul-Wadood, D. N. (2024). Investigating methods to identify and mitigate biases in AI algorithms to ensure fair and ethical outcomes. *2024 International Conference on Augmented Reality, Intelligent Systems, and Industrial Automation, ARIIA 2024*. https://doi.org/10.1109/ARIIA63345.2024.11051905

Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. *Journal of Computational Social Science*, *8*(2). https://doi.org/10.1007/s42001-025-00376-w

Rickman, S. (2025). Evaluating gender bias in large language models in long-term care. *BMC Medical Informatics and Decision Making*, *25*(1). https://doi.org/10.1186/s12911-025-03118-0

Romero-Arjona, M., Parejo, J. A., Alonso, J. C., Sánchez, A. B., Arrieta, A., & Segura, S. (2026). Meta-Fair: AI-assisted fairness testing of large language models. *Information and Software Technology*, *194*, 108075. https://doi.org/10.1016/j.infsof.2026.108075

Santagata, L., & de Nobili, C. (2025). More is more: Addition bias in large language models. *Computers in Human Behavior: Artificial Humans*, *3*, 100129. https://doi.org/10.1016/j.chbah.2025.100129

Sarrionandia, B., Peña-Fernández, S., & Pérez-Dasilva, J.-Á. (2025). Inteligencia artificial generativa y medios locales: análisis del sesgo territorial en cinco modelos de lenguaje. *Hipertext.Net*, (31), 201–213. https://doi.org/10.31009/hipertext.net.2025.i31.15

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, *63*(12), 54–63. https://doi.org/10.1145/3381831

Selten, F., Robeer, M., & Grimmelikhuijsen, S. (2023). 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Administration Review*, *83*(2), 263–278. https://doi.org/10.1111/puar.13602

Shahnejat Bushehri, A., Amirnia, A., Belkhiri, A., Keivanpour, S., de Magalhaes, F. G., & Nicolescu, G. (2024). Deep Learning-Driven Anomaly Detection for Green IoT Edge Networks. *IEEE Transactions on Green Communications and Networking*, *8*(1), 498–513. https://doi.org/10.1109/TGCN.2023.3335342

Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Social Science and Medicine*, *296*. https://doi.org/10.1016/j.socscimed.2022.114782

Skaiky, A. ali, Ali, H. M. S., Mohammed, A., & Mahdi, Z. A. (2025). Comprehensive Bias Mitigation in AI: Evaluating Pre-Processing, In-Processing, and Post-Processing Techniques for Fair Decision-Making. *2025 IEEE 4th International Conference on Computing and Machine Intelligence, ICMI 2025 - Proceedings*. https://doi.org/10.1109/ICMI65310.2025.11141086

Son, H., Lee, S., Kim, J., Park, H., Hwang, M. H., & Yi, G. S. (2024). BASE: a web service for providing compound-protein binding affinity prediction datasets with reduced similarity bias. *BMC Bioinformatics*, *25*(1), 1–26. https://doi.org/10.1186/S12859-024-05968-3

Stanley, E. A. M., Tsang, R. Y., Gillett, H., Souza, R., Vigneshwaran, V., Kang, C., McCradden, M. D., Wilms, M., & Forkert, N. D. (2026). Connecting algorithmic fairness and fair outcomes in a sociotechnical simulation case study of AI-assisted healthcare. *Nature Communications* . https://doi.org/10.1038/s41467-025-67470-5

Taisiq, S. A. (2026). CAS-NAS: A carbon-aware neural architecture search framework for sustainable AI development. *Engineering Science and Technology, an International Journal*, *76*. https://doi.org/10.1016/J.JESTCH.2026.102313

Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., Matsui, Y., Nozaki, T., Nakaura, T., Fujima, N., Tatsugami, F., Yanagawa, M., Hirata, K., Yamada, A., Tsuboyama, T., Kawamura, M., Fujioka, T., & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, *42*(1), 3–15. https://doi.org/10.1007/s11604-023-01474-3

Ueda, D., Walston, S. L., Fujita, S., Fushimi, Y., Tsuboyama, T., Kamagata, K., Yamada, A., Yanagawa, M., Ito, R., Fujima, N., Kawamura, M., Nakaura, T., Matsui, Y., Tatsugami, F., Fujioka, T., Nozaki, T., Hirata, K., & Naganawa, S. (2024). Climate change and artificial intelligence in healthcare: Review and recommendations towards a sustainable future. *Diagnostic and Interventional Imaging*, *105*(11), 453–459. https://doi.org/10.1016/j.diii.2024.06.002

Vajpayee, A. S., & Khobragade, D. (2024). The Problem Of Data Bias In Healthcare AI. *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare,*

Education and Industry, IDICAIEI 2024. https://doi.org/10.1109/IDICAIEI61867.2024.10842779

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(4). https://doi.org/10.1002/widm.1507

Wang, Q., Li, Y., & Li, R. (2024). Ecological footprints, carbon emissions, and energy transitions: the impact of artificial intelligence (AI). *Humanities and Social Sciences Communications*, *11*(1). https://doi.org/10.1057/s41599-024-03520-5

Wang, Y., Zhang, R., Yang, Q., Zhou, Q., Zhang, S., Fan, Y., Huang, L., Li, K., & Zhou, F. (2024). FairCare: Adversarial training of a heterogeneous graph neural network with attention mechanism to learn fair representations of electronic health records. *Information Processing and Management*, *61*(3). https://doi.org/10.1016/J.IPM.2024.103682

Wenzelburger, G., König, P. D., Felfeli, J., & Achtziger, A. (2024). Algorithms in the public sector. Why context matters. *Public Administration*, *102*(1), 40–60. https://doi.org/10.1111/padm.12901

Williamson, S. M., & Prybutok, V. (2024). Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. *Applied Sciences (Switzerland)*, *14*(2). https://doi.org/10.3390/app14020675

Xiang, A. (2026). Diagnosing the bias iceberg in large language models: A three-level framework of explicit, evaluative, and implicit gender bias. *Information Processing and Management*, *63*(3). https://doi.org/10.1016/j.ipm.2025.104554

Xiang, S., Xiang, Y., Lu, Y., Guo, Y., Tan, Z., & Wang, Y. (2023). Modeling and Optimization of Data Center Energy Consumption. *Proceedings - 2023 Panda Forum on Power and Energy, PandaFPE 2023*, 544–549. https://doi.org/10.1109/PandaFPE57779.2023.10140375

Yan, M., Mou, L., & Chen, L. (2026). Digital innovation and legal protection in China's predictive policing. *Computer Law & Security Review*, *60*, 106273. https://doi.org/10.1016/J.CLSR.2026.106273

Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, *2*. https://doi.org/10.1016/j.caeai.2021.100008

Yang, X., Li, Q., Qian, C., Wang, H., Wu, Y., & Wang, W. (2026). Bias Correction and Explainability Framework for Large Language Models: A Knowledge-Driven Approach. *Big Data and Cognitive Computing*, *10*(2). https://doi.org/10.3390/bdcc10020058

Yigitcanlar, T., Mehmood, R., & Corchado, J. M. (2021). Green artificial intelligence: towards an efficient, sustainable and equitable technology for smart cities and futures. *Sustainability (Switzerland)*, *13*(16). https://doi.org/10.3390/su13168952

Yin, Z., Wang, Z., Palikhe, A., & Zhang, W. (2026). Fairness Definitions in Language Models Explained. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *16*(1). https://doi.org/10.1002/widm.70063

Zhang, K. (2024). Promoting Equity: Assessing Algorithmic Fairness in Machine Learning Approaches for Predicting Alzheimer's Disease. *Proceedings - 2024 IEEE International Conference on Future Machine Learning and Data Science, FMLDS 2024*, 315–318. https://doi.org/10.1109/FMLDS63805.2024.00063

ANEXOS

1. Tabla de cumplimiento de objetivos específicos

Objetivo Específico	Secciones que evidencian su cumplimiento	Resumen del contenido y evidencia
OE1 y OE2: Sesgos cognitivos y datos de entrenamiento	**Sección 4.1** (Tipología y Patrones de Sesgos en la IAG)	Esta sección integra ambos objetivos al demostrar que los modelos no solo heredan sesgos de los **datasets no representativos** (OE2), sino que también replican **patrones de razonamiento sesgado** como el de anclaje y el de adición (OE1). Se destaca el concepto del **"iceberg del sesgo"**, donde los sesgos implícitos en las capas profundas son hasta 12.3 veces mayores que los visibles.
OE3: Casos documentados	**Sección 4.2** (Casos Documentados por Sector)	Se detallan impactos críticos en **salud** (diagnóstico por imagen y registros electrónicos) y **justicia** (sistema COMPAS y policía predictiva). La evidencia muestra cómo variables aparentemente neutras actúan como *proxies* de raza o nivel socioeconómico, perpetuando discriminaciones estructurales.
OE4: Estrategias de mitigación	**Sección 4.3** (Estrategias de Mitigación Evaluadas)	Evalúa técnicas de **pre, en y post-procesamiento**, así como marcos de gobernanza como **NoBIAS** y el **Reglamento Europeo de IA (AI Act)**. Se documenta el "trade-off" donde mejorar la equidad puede

		reducir la precisión del modelo hasta en 10 puntos porcentuales.
OE5: Impacto ambiental	**Sección 4.4** (Impacto Ambiental de la IAG)	Provee evidencia cuantitativa sobre el **consumo de GPT-3 (1.287 MWh)** y las emisiones de CO2. Analiza la efectividad de la **IA Verde**, reportando que técnicas como la poda de modelos y la cuantización pueden reducir el consumo energético entre un 13% y un 115%.

2. Ficha Tecnica

ESPACIO PARA SER DILIGENCIADO POR EL ESTUDIANTE
Título de la Propuesta:
Inteligencia Artificial Generativa: ética, sesgos y sostenibilidad
Opción de grado:
☐ Proyecto de investigación ☒Monografía ☐ Desarrollo tecnológico ☐ Pasantía empresarial
Objetivo General
Examinar la influencia de los sesgos cognitivos, la calidad de los datos de entrenamiento y el impacto ambiental en el desarrollo de modelos de Inteligencia Artificial Generativa, mediante una revisión sistemática de literatura científica de alto impacto, con el propósito de valorar sus implicaciones éticas y las estrategias vigentes para su mitigación.
Objetivos específicos
1. Explorar cómo los sesgos cognitivos y las bias influyen en el desarrollo y comportamiento de modelos de Inteligencia Artificial Generativa, a partir de una revisión teórica de enfoques conceptuales y técnicos relacionados. 2. Investigar el papel de los datos de entrenamiento en la introducción de sesgos involuntarios y sus consecuencias éticas, a través del análisis de artículos indexados que documenten ejemplos y estudios de caso. 3. Analizar casos documentados en los que modelos de Inteligencia Artificial han mostrado sesgos significativos, y evaluar sus efectos en sectores como salud, justicia y redes sociales. 4. Evaluar estrategias actuales para mitigar los sesgos en modelos de Inteligencia Artificial Generativa, considerando enfoques técnicos y éticos reportados en literatura científica y libros especializados. 5. Determinar el impacto ambiental asociado al entrenamiento y operación de modelos de Inteligencia Artificial Generativa, mediante el estudio de métricas como consumo energético, reportadas por organizaciones como Google AI, OpenAI o informes técnicos.

ESPACIO PARA SER DILIGENCIADO POR EL ESTUDIANTE		
Autores		
Nombres y Apellidos	**Correo Electrónico**	**Teléfono**
John Jairo Vergel Camacho	John.vergel@usantoto.edu.co	3134835436
Juan Felipe López González	juanf.lopezg@usantoto.edu.co	3106138259
Director	**Línea de Investigación del semillero o temática asociada**	
Diego Alejandro Vela Beltran	Ingeniería del Software ☒ Software Educativo ☐	