Revista Jurídica de la Universidad Americana

https://doi.org/10.30545/juridica Vol. 14 N°1 (enero-junio) 2026. p. 61-76.

ARTÍCULO ORIGINAL

https://doi.org/10.30545/juridica.2026.ene-jun.5

Inteligencia Artificial y Sesgos de Género en Decisiones Judiciales Brasileñas

Artificial Intelligence and Gender Bias in Brazilian Court Decisions

Camila Henning Salmoria[1]

[1] Tribunal de Justicia del Paraná, Brasil.

RESUMEN

El presente artículo analiza el empleo de modelos de inteligencia artificial generativa para detectar y corregir sesgos de género en resoluciones judiciales brasileñas, tomando como marco el Protocolo para el Juicio con Perspectiva de Género del CNJ. Mediante un enfoque cualitativo y exploratorio, se aplicaron técnicas de Procesamiento de Lenguaje Natural en un estudio de caso múltiple que evaluó tres configuraciones de instrucciones o prompts (ML, CL y MCL). Los resultados evidencian que la calibración normativa y semántica de las instrucciones resulta fundamental para la precisión de las inferencias. El modelo híbrido MCL demostró ser el más eficaz en la identificación de estereotipos, la culpabilización de la víctima y la omisión de situaciones de vulnerabilidad. Se concluye que la inteligencia artificial generativa puede ejercer funciones formativas y de auditoría institucional, siempre que opere bajo parámetros normativos rigurosos y supervisión humana continua. Su implementación exige la expansión progresiva del corpus analizado y una mayor validación empírica para su eventual consolidación en la praxis judicial.

Palabras clave: Inteligencia artificial generativa, sesgos de género, resoluciones judiciales, PLN, igualdad sustantiva.

[1] **Correspondencia:** camilahsalmoria@gmail.com
Conflicto de Interés: Ninguno.
Financiamiento: Ninguna.
Recibido: 15/11/2025; aprobado: 30/03/2026; publicado: 26/06/2026.

Sitio web: https://revistacientifica.uamericana.edu.py/index.php/revistajuridicaua/index

ABSTRACT
This article analyzes the use of generative artificial intelligence models to detect and correct gender bias in Brazilian judicial decisions, based on the CNJ's Protocol for Judgment with a Gender Perspective. Adopting a qualitative and exploratory approach, Natural Language Processing techniques were applied in a multiple-case study that evaluated three prompt configurations (ML, CL, and MCL). The results indicate that normative and semantic calibration of instructions is essential for the quality of the inferences. The hybrid MCL model proved to be the most effective in identifying stereotypes, victim-blaming, and omissions regarding vulnerability. The findings suggest that generative AI can perform formative and institutional auditing functions, provided it is calibrated according to clear normative parameters and subject to continuous human oversight. Its application requires the progressive expansion of the corpus and further empirical validation for its eventual consolidation in judicial practice.

Keywords: Generative artificial intelligence, gender bias, judicial decisions, NLP, substantive equality.

INTRODUCCIÓN

La inteligencia artificial generativa se ha ido incorporando en el Poder Judicial como una herramienta para optimizar la redacción jurídica, abarcando desde la corrección gramatical y ortográfica hasta el perfeccionamiento del discurso. Estudios a nivel internacional evidencian que los modelos de lenguaje de gran escala (LLM, por sus siglas en inglés) pueden emplearse eficazmente para detectar sesgos de género en las resoluciones judiciales. En este contexto, el presente trabajo analiza la aplicación de dichos modelos como instrumentos para la detección y corrección de sesgos de género en decisiones judiciales brasileñas, a la luz de las directrices del Protocolo para el Juicio con Perspectiva de Género del Consejo Nacional de Justicia (Conselho Nacional de Justiça [CNJ], 2021).

La investigación parte de la siguiente interrogante: ¿en qué medida los modelos de inteligencia artificial generativa, calibrados mediante parámetros normativos, pueden identificar y corregir sesgos de género en textos decisorios del Poder Judicial brasileño? La hipótesis orientadora sostiene que, al ser guiados por instrucciones ético-jurídicas, estos modelos poseen la capacidad semántica y contextual necesaria para reconocer patrones discursivos discriminatorios y proponer reescrituras que sean compatibles con los principios de igualdad sustantiva y no discriminación. De este modo, la IA tiene el potencial de actuar como una herramienta de apoyo, tanto para la auditoría institucional como para la formación continua de la magistratura, promoviendo un ejercicio jurisdiccional más consciente y comprometido con los derechos humanos.

El objetivo general del estudio consiste en evaluar la capacidad de los modelos de inteligencia artificial generativa, calibrados mediante parámetros normativos derivados del Protocolo del CNJ, para identificar y sugerir correcciones de sesgos de género en sentencias brasileñas.

Como objetivos específicos, se proponen:

- Verificar la aptitud del modelo para detectar patrones discursivos de sesgo conforme a las categorías establecidas en el Protocolo.
- Comparar el desempeño metodológico de tres configuraciones distintas de *prompts* (ML, CL y MCL).
- Examinar el potencial formativo e institucional derivado del uso experimental de estos modelos en contextos de auditoría judicial.

La relevancia científica y social de este trabajo radica en la brecha que aún existe entre el desarrollo técnico de la IA y su aplicación ética dentro del sistema de justicia. A pesar de los avances tecnológicos, persiste una escasez de estudios empíricos que sometan a prueba la compatibilidad de estos sistemas con los marcos normativos de género, especialmente en contextos decisorios. Por consiguiente, la investigación busca aportar evidencia sobre la viabilidad de un uso responsable de la IA generativa como instrumento de perfeccionamiento institucional, fortaleciendo así las políticas públicas de equidad y contribuyendo a la consolidación de una cultura jurídica sensible a las desigualdades estructurales.

METODOLOGÍA

El trabajo se desarrolló bajo un enfoque cualitativo de carácter exploratorio, orientado a la aplicación de técnicas de Procesamiento de Lenguaje Natural (PLN) y de inteligencia artificial generativa (IAG) en el contexto del Poder Judicial brasileño. El punto de partida fue la premisa de que los sistemas comerciales de IA, al ser calibrados con parámetros normativos claros, pueden utilizarse para detectar sesgos de género en textos judiciales y sugerir reescrituras alineadas con el Protocolo del CNJ (2021). Se asumió que los modelos de lenguaje de gran escala cuentan con la capacidad semántica y contextual suficiente para reconocer patrones discursivos asociados a desigualdades de género, siempre que operen bajo instrucciones que prioricen principios éticos, jurídicos y de derechos humanos.

En consonancia con los objetivos planteados, la estrategia metodológica consistió en diseñar y ejecutar un procedimiento experimental comparativo para evaluar el desempeño de tres configuraciones de *prompts* (ML, CL y MCL, uyos textos íntegros constan en el Apéndice 1), basándose en criterios cualitativos de detección, estructuración analítica y adherencia normativa al Protocolo. Así, la metodología no introduce un objetivo sustantivo nuevo, sino que operacionaliza de forma empírica el objetivo general mediante un diseño de estudio de caso múltiple con aplicación experimental controlada.

Diseño del Estudio

El estudio se concibió bajo el modelo de investigación aplicada con estudio de caso múltiple, inspirado en el diseño propuesto por Yin (2005). Esta elección permitió conjugar la observación de fenómenos concretos —las manifestaciones de sesgo en resoluciones reales— con el examen conceptual de las bases del juicio con perspectiva de género.

La construcción del diseño empírico estuvo precedida por una revisión sistemática de la literatura internacional para mapear el estado del arte sobre el uso de la IA en el análisis de sesgos judiciales. Se identificaron investigaciones en Estados Unidos, Portugal, Fiji y Brasil, las cuales confirmaron la pertinencia de la temática y brindaron elementos para la

adaptación metodológica al contexto nacional. En particular, sirvieron de fundamento técnico y ético:

- El estudio de Chen et al. (2020) sobre sesgos de género en tribunales norteamericanos.
- Las investigaciones de Pinto et al. (2020; 2021) referidas al uso del PLN para detectar discriminación en decisiones portuguesas.
- La aplicación de técnicas de PLN en sentencias penales de Fiji (Sexton & Tozzi, 2020).
- El experimento brasileño de Benatti (2023), aplicado a sentencias civiles.

Contexto y Corpus de la Investigación

El corpus empírico se conformó por resoluciones judiciales brasileñas seleccionadas por presentar contextos donde la perspectiva de género resultaba medular. Los textos se extrajeron de la base consolidada en la obra *Reescrevendo Decisões com Perspectiva de Gênero*, organizada por Severi (2023), que compila casos emblemáticos de argumentación sesgada e ignorancia de la vulnerabilidad de la víctima. Esta selección garantizó el trabajo con material previamente validado en términos de relevancia temática y presencia de sesgos discursivos.

Se seleccionaron cinco casos de la mencionada obra, la cual es de acceso abierto (Severi, 2023). Los criterios operativos de selección fueron: (i) la presencia explícita de controversias sobre violencia sexual o vulnerabilidad femenina; y (ii) la existencia de un debate probatorio susceptible de revelar patrones discursivos sesgados.

El agente conversacional fue configurado experimentalmente por la autora a través de la ingeniería de *prompts*, sin un entrenamiento adicional del modelo base y con fines estrictamente investigativos. Dicho agente se programó para identificar estereotipos, distorsiones interpretativas u omisiones de género en cada resolución. Asimismo, fue instruido para clasificar los fragmentos y proponer reescrituras acordes al Protocolo del CNJ (2021) bajo las siguientes categorías: uso de estereotipos de género, culpabilización de la víctima, desconsideración de la palabra de la víctima, aplicación neutral de conceptos jurídicos y omisión de factores interseccionales. Estas categorías orientaron la fase de codificación y el análisis cualitativo.

Variables y Procedimientos

Para la ejecución de las pruebas, se definieron variables dependientes e independientes con el fin de permitir un análisis estructurado. Las variables dependientes incluyeron la capacidad de la IA para detectar sesgos conforme al Protocolo del CNJ (2021), la coherencia argumentativa de las justificaciones y la precisión de las sugerencias de reescritura. Por su parte, las variables independientes fueron el nivel de claridad textual de las decisiones y el grado de detalle de las instrucciones contenidas en los *prompts*.

El proceso experimental se llevó a cabo en tres rondas sucesivas, cada una dedicada a una versión distinta de *prompt*. El *prompt* ML fue diseñado para promover un análisis interpretativo y didáctico, reproduciendo el razonamiento de un magistrado o una magistrada con formación en perspectiva de género. El *prompt* CL, en cambio, priorizó la objetividad técnica y la estructuración de las

respuestas en etapas fijas, privilegiando la identificación y la clasificación sistemática de los sesgos. A partir de la comparación entre estas dos versiones, se desarrolló el *prompt* MCL, que integró las virtudes analíticas y normativas de los anteriores, buscando un equilibrio entre el rigor jurídico y la claridad comunicativa.

Las respuestas generadas en cada ronda se sometieron a un análisis comparado basado en criterios de consistencia interna, exhaustividad argumentativa y apego a las directrices del Protocolo. Para garantizar la robustez interpretativa, se empleó una triangulación metodológica que combinó tres niveles de verificación: la revisión manual de las respuestas producidas por la IA, la comparación entre las distintas versiones de los *prompts* y la confrontación de los resultados con decisiones y notas técnicas oficiales del CNJ (2021).

Aspectos Técnicos y Éticos

Las pruebas se realizaron con el modelo comercial ChatGPT-4, desarrollado por OpenAI (2024), seleccionado por su elevada capacidad semántica, amplia ventana de *tokens* y estabilidad discursiva en la generación de respuestas complejas. La configuración del agente observó rigurosos principios éticos y de seguridad de la información, especialmente respecto a la integridad de los datos y la imparcialidad de las respuestas. Se incluyeron instrucciones explícitas para que el modelo utilizara un lenguaje neutro y respetuoso, evitara inferencias no fundamentadas y garantizara la estricta anonimización de los casos.

El procedimiento técnico se dividió en tres etapas interdependientes: preparación de los datos, ejecución de las pruebas y análisis crítico de los resultados. En la preparación, los textos fueron revisados y normalizados para asegurar su coherencia semántica; en la ejecución, cada decisión fue procesada íntegramente por el modelo, registrándose las respuestas obtenidas; en el análisis, las salidas fueron clasificadas conforme a las categorías de sesgo y evaluadas en cuanto a su adherencia normativa.

Declaración sobre el uso de Inteligencia Artificial

La presente investigación empleó herramientas de inteligencia artificial generativa con fines experimentales y bajo estricta supervisión humana, en conformidad con los principios éticos del Consejo Nacional de Justicia (CNJ) y de la UNESCO. Durante la redacción del manuscrito, dichas herramientas se utilizaron exclusivamente para labores de revisión gramatical, ortográfica y de traducción, sin interferir en el contenido intelectual, la interpretación de los resultados ni en el juicio analítico de la autora.

Limitaciones del Estudio

Como toda investigación cualitativa y exploratoria, el estudio presentó limitaciones inherentes a su diseño experimental. La muestra de decisiones analizadas fue restringida y no probabilística, priorizándose la demostración de la viabilidad metodológica en detrimento de la representatividad estadística.

La interpretación semántica de la IA mostró dependencia respecto a la calidad textual de las decisiones, la precisión de las instrucciones y la calibración de los *prompts*, especialmente cuando el discurso judicial presentaba ambigüedades o contradicciones internas. Asimismo, se constató la necesidad de perfeccionar el corpus jurídico disponible en lengua portuguesa.

Estas limitaciones no comprometen la consistencia de los resultados, pero refuerzan la importancia de la supervisión humana continua y de la reflexividad metodológica. La experiencia empírica indicó que la eficacia de la IA en el contexto judicial no depende únicamente de su capacidad técnica, sino principalmente de la claridad de las instrucciones, la calidad de los datos y la mediación ética que orienta su aplicación. La metodología delineada constituye, así, una base sólida para el desarrollo de futuras investigaciones y para el avance institucional en la creación de herramientas de IA orientadas a la identificación y mitigación de sesgos de género en el Poder Judicial brasileño.

RESULTADOS

Los resultados obtenidos en la etapa empírica y en los análisis comparados confirman la viabilidad técnica, metodológica y epistémica del uso de modelos de lenguaje basados en inteligencia artificial generativa (IAG) para la detección de sesgos de género en decisiones judiciales. El experimento sugirió que, cuando tales modelos están calibrados con parámetros normativos claros y anclados en marcos ético-jurídicos consistentes —especialmente el Protocolo para el Juicio con Perspectiva de Género del Consejo Nacional de Justicia (CNJ)—, son capaces de reconocer patrones discursivos sutiles relacionados con estereotipos de género, culpabilización de la víctima, minimización de la violencia y aplicación neutral de conceptos jurídicos.

La coherencia argumentativa y la fundamentación normativa presentes en las respuestas evidencian que la IAG puede emplearse no solo como instrumento de auditoría judicial, sino también como herramienta de formación y sensibilización de la magistratura. El análisis empírico indica que, ante instrucciones bien delimitadas, los sistemas generativos producen inferencias compatibles con los principios de igualdad sustantiva y no discriminación, señalando una convergencia favorable entre la tecnología y los compromisos internacionales asumidos por el Estado brasileño en materia de derechos humanos y justicia de género.

A efectos de sistematización comparativa, los principales criterios de desempeño de los *prompts* ML, CL y MCL se presentan en la tabla 1, la cual sintetiza el nivel de detección de sesgos, el grado de estructuración analítica, la explicabilidad y la adherencia normativa.

Tabla 1. Desempeño comparativo de los prompts ML, CL y MCL.

Criterio	ML	CL	MCL
Detección de estereotipos	Moderada	Alta	Alta
Estructura analítica	Baja	Alta	Alta
Explicabilidad	Alta	Moderada	Alta
Adherencia al Protocolo	Moderada	Alta	Muy Alta

Resultados Empíricos de las Pruebas con IA

Durante la fase experimental se realizaron pruebas sucesivas con tres variaciones de *prompts*, aquí denominadas *prompt* ML, *prompt* CL y *prompt* MCL. Cada configuración representó un nivel de evolución metodológica en la interacción entre el lenguaje natural y el razonamiento jurídico orientado por el Protocolo del CNJ.

El *prompt* ML, formulado inicialmente con énfasis en la función pedagógica e

interpretativa, orientó al modelo a actuar como un magistrado ficticio, reproduciendo la lógica deliberativa del juicio con perspectiva de género. Sus respuestas fueron predominantemente narrativas y explicativas, articulando normas internacionales con directrices nacionales, y priorizando la didáctica y el esclarecimiento conceptual sobre la estructura formal de la decisión. A pesar de su elevada calidad argumentativa, el ML presentaba menor sistematicidad en la categorización de los sesgos, lo que limitaba su replicabilidad en contextos de auditoría comparada.

Posteriormente, el *prompt* CL introdujo un avance sustancial. Formulado para actuar como un agente especializado en el análisis de sesgos de género, el modelo pasó a emplear una estructura analítica fija, segmentando la respuesta en cuatro elementos: fragmento identificado, clasificación del sesgo, explicación técnica y sugerencia de corrección. Esta estandarización amplió la precisión metodológica y favoreció la reproducibilidad de los resultados. El CL resultó especialmente eficaz en la detección de expresiones típicas de culpabilización de la víctima, particularmente en fragmentos que atribuían a la mujer comportamientos supuestamente "provocadores" o "negligentes", como el consumo de alcohol previo al crimen. En tales situaciones, el modelo identificó correctamente el sesgo y fundamentó su crítica basándose en la Parte II, ítem 7.d, del Protocolo del CNJ, proponiendo redacciones alternativas como: "El consumo de alcohol por parte de la víctima no puede ser considerado factor de imputación de culpa, so pena de violación del principio de no discriminación".

Finalmente, el *prompt* MCL —resultado de la fusión entre el ML y el CL— presentó el desempeño más avanzado y consistente. Este modelo híbrido conjugó el rigor conceptual y normativo del ML con la claridad estructural y la objetividad técnica del CL, alcanzando un notable equilibrio entre la densidad analítica y la precisión jurídica. El MCL hizo referencias expresas a secciones del Protocolo, demostrando una alta adherencia metodológica y dominio de la hermenéutica normativa. Gracias a ello, fue capaz de interpretar fragmentos decisorios con sensibilidad hacia las dinámicas de poder y la desigualdad de género, reconociendo la invisibilización de la vulnerabilidad de la víctima y la aplicación indebida de conceptos jurídicos aparentemente neutros.

El análisis comparado indicó que la calibración de los *prompts* fue determinante para el desempeño del modelo, confirmando la hipótesis de que la ingeniería de *prompts* constituye una etapa esencial de la regulación ética y técnica de la IAG aplicada al ámbito judicial. Las versiones más detalladas y normativamente contextualizadas, especialmente el MCL, produjeron respuestas con mayor consistencia lógica, densidad argumentativa y conformidad jurídica.

Síntesis Interpretativa

En relación con el primer objetivo específico —verificar la capacidad de detección de sesgos conforme al Protocolo—, los resultados evidenciaron que el modelo híbrido MCL presentó la mayor precisión en la identificación de estereotipos, culpabilización de la víctima y omisiones interseccionales.

Respecto al segundo objetivo —comparar el desempeño metodológico de los *prompts*—, se

constató que la calibración normativa influye directamente en la densidad argumentativa y en la coherencia hermenéutica de las respuestas.

Finalmente, en cuanto al tercer objetivo —examinar el potencial formativo e institucional—, el uso experimental de la herramienta reveló su probada capacidad para funcionar como un instrumento reflexivo de apoyo a la capacitación judicial.

La elaboración de *prompts* orientados por epistemologías feministas del derecho y por marcos normativos oficiales resultó decisiva para alcanzar resultados éticamente satisfactorios y jurídicamente válidos. En otras palabras, la calibración del *prompt* opera, en sí misma, como una forma de regulación algorítmica, siendo la intervención humana indispensable para garantizar la integridad hermenéutica y los límites éticos del uso de la IA en el sistema de justicia.

Los resultados de la investigación empírica indican que el uso de modelos de lenguaje basados en IAG puede representar una inflexión relevante en la formación de la magistratura, así como en la auditoría institucional de las decisiones judiciales. Al detectar sesgos de género de forma estructurada e interpretable, los modelos evaluados —en especial el MCL— evidenciaron un gran potencial pedagógico y operativo para reforzar la aplicación del Protocolo para el Juicio con Perspectiva de Género en la práctica jurisdiccional cotidiana.

En el ámbito formativo, el uso de la IA se mostró eficaz como instrumento de sensibilización y autoevaluación. El análisis automatizado de decisiones, acompañado de explicaciones contextualizadas sobre los sesgos identificados, permite que los juzgadores reconozcan cómo determinadas elecciones lingüísticas y narrativas pueden reproducir desigualdades de género, incluso en ausencia de intención discriminatoria. Esta retroalimentación inmediata contribuye a la internalización de las directrices del Protocolo y al perfeccionamiento continuo de la argumentación judicial. De este modo, la tecnología actúa como un espejo hermenéutico: devuelve al juzgador una lectura crítica de su propio discurso y refuerza la dimensión reflexiva del acto de juzgar.

La experiencia también reveló que la implementación de estos agentes de análisis puede favorecer el desarrollo de programas de capacitación basados en estudios de casos simulados, en los cuales decisiones reales son analizadas por el sistema y discutidas colectivamente. Esta metodología, centrada en la interacción entre humanos y máquinas, estimula el aprendizaje activo, convierte el juicio con perspectiva de género en una práctica experiencial y promueve la convergencia entre el razonamiento jurídico y la conciencia social. En cursos de formación inicial o continua, la IA puede actuar como mediadora, ofreciendo retroalimentación objetiva y estandarizada sobre la adecuación de las decisiones a las directrices del CNJ.

En el plano institucional, los hallazgos indican que la IA generativa puede emplearse como una herramienta de auditoría algorítmica interna, capaz de auxiliar en la verificación sistemática de la conformidad de las decisiones con los principios de igualdad y no discriminación. La aplicación de estos modelos en muestras de sentencias o *acórdãos* (resoluciones de tribunales colegiados) permite identificar patrones de lenguaje que señalen la reproducción de estereotipos, el desequilibrio en la valoración probatoria o la omisión de marcadores de vulnerabilidad. Tales hallazgos

pueden sustentar informes periódicos sobre la aplicación del Protocolo y apoyar políticas de gobernanza judicial orientadas por la evidencia.

Esta función de auditoría preventiva presenta beneficios adicionales: reduce el riesgo de sentencias discriminatorias, amplía la transparencia institucional y fortalece la legitimidad pública del sistema de justicia. Al visibilizar las estructuras discursivas que sostienen desigualdades históricas, la tecnología contribuye a la consolidación de una cultura organizacional comprometida con la equidad y la imparcialidad sustantiva.

Finalmente, la investigación corroboró que la eficacia de los modelos depende ineludiblemente de su integración con una supervisión humana calificada. El uso de la IA debe concebirse como un apoyo metodológico y no como un sustituto del razonamiento judicial. La amalgama entre el análisis algorítmico y la interpretación crítica garantiza que las recomendaciones generadas sean comprendidas, validadas y, cuando sea necesario, reinterpretadas a la luz del contexto fáctico y jurídico. De esta forma, la tecnología cumple una función de apoyo reflexivo a la decisión, asegurando que el juicio con perspectiva de género prevalezca como un acto humano, consciente y socialmente responsable.

En síntesis, la aplicación experimental de los *prompts* ML, CL y MCL demostró que la inteligencia artificial puede operar simultáneamente como herramienta de formación y de auditoría, mejorando las condiciones institucionales para el cumplimiento efectivo del Protocolo para el Juicio con Perspectiva de Género y promoviendo una cultura de autocrítica y transparencia en el ejercicio de la jurisdicción.

DISCUSIÓN

Los resultados obtenidos sugieren que la inteligencia artificial generativa puede convertirse en un instrumento cardinal para la promoción de una justicia más equitativa, siempre que su utilización sea guiada por principios éticos y normativos, y supervisada por agentes humanos calificados. La capacidad demostrada por la IA para identificar patrones lingüísticos y argumentativos asociados a sesgos de género revela no solo un avance tecnológico, sino una nueva vía para el fortalecimiento del compromiso institucional con la igualdad sustantiva. Esta constatación amplía el alcance tradicional del análisis jurídico, el cual pasa a incorporar metodologías basadas en datos y evidencia discursiva como un medio para perfeccionar la calidad y la imparcialidad de la decisión judicial.

Interpretación de los Resultados a la Luz de los Objetivos y Limitaciones

El análisis de los resultados evidencia que el desempeño del agente de IA varió en función de la estructura de los *prompts* y del nivel de contextualización jurídica proporcionado. Esta variación refuerza las conclusiones de estudios recientes sobre la ingeniería de *prompts*, según los cuales el diseño de las instrucciones es un factor determinante para la calidad de la respuesta y para la mitigación de los sesgos residuales del propio modelo. El rendimiento satisfactorio del sistema en la identificación de pasajes de culpabilización de la víctima y de neutralidad aparente confirma que, cuando está adecuadamente orientada por directrices normativas, la tecnología es capaz de captar matices discursivos y de reproducir razonamientos jurídicos coherentes con los

principios de género y de igualdad material consagrados en el Protocolo del CNJ.

No obstante, el carácter exploratorio de la investigación y la limitación del corpus analizado imponen cautela respecto a la generalización de las conclusiones. La diversidad estilística y lingüística de las resoluciones judiciales brasileñas, sumada a las diferencias regionales y a las especificidades de cada rama de la Justicia, sugiere la necesidad de ampliar el espectro empírico y de profundizar en el proceso de validación de las respuestas. En este sentido, la investigación opera como una prueba de concepto: demuestra la viabilidad técnica y metodológica de aplicar la IA al ámbito de la imparcialidad judicial, pero al mismo tiempo advierte sobre la necesidad de un proceso de maduración más profundo, tanto a nivel institucional como científico, para su consolidación como herramienta de uso regular y confiable.

Diálogo con las Investigaciones Internacionales

La investigación empírica y comparativa arrojó tres hallazgos centrales de gran relevancia teórica y metodológica para el campo de la justicia algorítmica.

El primer hallazgo se refiere a la viabilidad técnica y epistemológica de la aplicación de modelos de lenguaje generativo para la detección de sesgos de género en decisiones judiciales. Cuando se instruye mediante *prompts* calibrados normativamente —en especial el MCL, que amalgama las virtudes estructurales del CL con la adherencia hermenéutica del ML—, la inteligencia artificial demuestra una capacidad robusta para identificar y clasificar patrones discursivos complejos asociados a la reproducción de estereotipos de género, a la culpabilización de la víctima y a la aplicación neutral de conceptos jurídicos. Esta constatación está en plena consonancia con las conclusiones de Chen et al. (2020), cuya investigación empírica en más de 380 mil decisiones estadounidenses demostró que los magistrados que reproducen estereotipos de género en sus textos tienden a dictar resoluciones más restrictivas para los derechos de las mujeres. En la misma línea, el estudio portugués de Pinto et al. (2020; 2021) y la experiencia brasileña de Benatti (2023) refuerzan la eficacia del procesamiento de lenguaje natural (PLN) para detectar patrones lingüísticos discriminatorios, validando el uso de arquitecturas neuronales y *word embeddings* como instrumentos de diagnóstico de sesgos implícitos.

El segundo hallazgo indica que los sesgos más recurrentes en las decisiones analizadas son de naturaleza interpretativa y estructural, vinculados estrechamente a la valoración de la prueba, a la construcción narrativa de los hechos y a la presunta neutralidad argumentativa —un fenómeno profusamente documentado por la literatura internacional (Chen et al., 2020; Salmoria, 2024; Sexton & Tozzi, 2020). Este tipo de sesgo no suele manifestarse mediante expresiones abiertamente discriminatorias, sino a través de razonamientos jurídicos que naturalizan jerarquías de género y reproducen desigualdades sistémicas, tales como la sistemática desconfianza respecto a la palabra de la víctima o la exigencia de resistencia física para la caracterización de la violencia sexual. El desempeño del modelo MCL, que fundamentó sus respuestas basándose en secciones específicas del Protocolo para el Juicio con Perspectiva de Género del CNJ (2021),

demostró que la IA puede reconocer estas asimetrías discursivas, señalando no solo el error argumentativo, sino también la matriz estructural del prejuicio subyacente —una competencia interpretativa que aproxima de manera notable el análisis algorítmico al razonamiento jurídico-humanista.

El tercer hallazgo concierne a la explicabilidad y a la legitimidad institucional de las respuestas generadas por la IA, factores que resultan determinantes para su aceptación en el entorno jurídico. Como subrayan Gabriel (2022) y Coeckelbergh (2023), la transparencia interpretativa constituye un prerrequisito ético y jurídico ineludible para que los sistemas automatizados puedan integrarse de manera legítima a la práctica judicial. En el experimento realizado, se constató que las respuestas explicables —es decir, aquellas que explicitaban la secuencia lógica de su razonamiento, citaban secciones pertinentes del Protocolo y ofrecían justificaciones fundamentadas para sus conclusiones— lograban satisfacer cabalmente estos criterios. Esto ratifica que la inteligibilidad de la decisión algorítmica es una condición *sine qua non* para su legitimidad democrática (Peixoto & Bonat, 2023), puesto que únicamente la explicación comprensible habilita el control humano y la consecuente responsabilización institucional.

Entre las limitaciones identificadas, sobresalen tres aspectos interdependientes. El primero es la ausencia de un corpus jurídico extenso y debidamente depurado, lo que restringe las posibilidades de entrenamiento de modelos en portugués y reduce la precisión semántica de determinadas detecciones (Benatti, 2023; Pinto et al., 2020). El segundo radica en la dependencia respecto a la calidad textual de las decisiones judiciales objeto de análisis, las cuales frecuentemente presentan inconsistencias de redacción y variaciones terminológicas que suponen un reto para los algoritmos de PLN, dificultando así la normalización lingüística y la identificación de patrones sutiles (Carmo et al., 2023). Finalmente, la necesidad de una calibración continua de los *prompts* constituye un desafío metodológico inherente al empleo de modelos generativos. Tal como se constató en la fase experimental, ligeras modificaciones en la formulación del *prompt* pueden alterar sustancialmente la densidad normativa y la coherencia interpretativa de la respuesta. Este hallazgo respalda la tesis de que la ingeniería de *prompts* representa una forma emergente de regulación algorítmica (Boden, 2020; Salmoria, 2024).

Estas limitaciones, sin embargo, no invalidan los resultados obtenidos; más bien, revelan la naturaleza exploratoria de la investigación y el potencial para el desarrollo de una infraestructura algorítmica con orientación jurídica. Tal como observaron Chen et al. (2020) y Pinto et al. (2021), la madurez técnica de los sistemas de IA depende en menor medida de la cantidad de datos y en mayor medida de la calidad de la curaduría y del anclaje normativo.

El presente trabajo concuerda con estudios previos al reconocer la centralidad del lenguaje en la construcción de las decisiones judiciales, pero se distingue por emplear modelos generativos capaces no solo de analizar, sino también de reformular discursivamente los textos. Esta característica le otorga al estudio un carácter innovador, dado que la inteligencia artificial no se limita a un mero análisis descriptivo de los sesgos, sino que actúa también de forma prescriptiva, sugiriendo alternativas de redacción compatibles con los

principios de igualdad y no discriminación. Al proponer el uso de una tecnología que detecta y corrige simultáneamente las distorsiones discursivas, la presente investigación contribuye al campo de la justicia algorítmica desde una perspectiva emancipatoria, en estricta consonancia con la *Recomendación sobre la Ética de la Inteligencia Artificial* de la UNESCO (2021), la cual promueve el uso de esta tecnología como un vector de inclusión y de justicia social.

CONCLUSIONES

Los resultados obtenidos confirmaron la viabilidad técnica y metodológica del uso de modelos de lenguaje basados en inteligencia artificial generativa como instrumentos auxiliares para la detección y corrección de sesgos de género en las decisiones judiciales brasileñas. La aplicación de los *prompts* ML, CL y MCL evienció que la calibración normativa y semántica es determinante para la consistencia y la calidad de las respuestas producidas. Entre los tres modelos evaluados, el MCL presentó el mejor desempeño, revelando una mayor adherencia a las directrices del Protocolo para el Juicio con Perspectiva de Género del Consejo Nacional de Justicia (CNJ, 2021), así como una superior densidad argumentativa y coherencia jurídica. De este modo, el estudio cumplió con el objetivo propuesto de evaluar la capacidad de la IA generativa para actuar como herramienta formativa y de auditoría discursiva, contribuyendo a prácticas jurisdiccionales más sensibles a la igualdad sustantiva.

La evidencia empírica obtenida resulta consistente con la hipótesis formulada, en la medida en que los modelos calibrados normativamente mostraron capacidad para identificar y sugerir correcciones a patrones discursivos discriminatorios, dentro de los límites del diseño exploratorio adoptado. Los hallazgos empíricos indicaron que la eficacia de la tecnología no reside únicamente en su capacidad estadística, sino principalmente en la sofisticación semántica y ética de su configuración inicial. La ingeniería de *prompts* se reveló como un componente esencial de la regulación algorítmica y un factor decisivo para obtener resultados interpretables, jurídicamente válidos y éticamente aceptables.

El análisis integrado de las secciones permitió constatar que la fundamentación teórica proporcionó el marco necesario para situar la discusión sobre la IA en el contexto de la justicia de género, mientras que la metodología —basada en el estudio de caso múltiple y en técnicas de procesamiento de lenguaje natural— consolidó la credibilidad de los resultados. Las pruebas empíricas demostraron, de manera comparativa, cómo la evolución de los *prompts* elevó el nivel de precisión analítica y de adherencia a las normas del CNJ, culminando en la elaboración de un modelo híbrido capaz de articular la claridad técnica con la sensibilidad social. La síntesis interpretativa final reforzó que la calibración semántica de la IA constituye, en sí misma, una forma de regulación, en la medida en que define los límites y las potencialidades de su intervención en el campo jurídico.

Desde un punto de vista científico e institucional, la investigación contribuye al debate contemporáneo sobre la ética y la gobernanza de la IA en el Poder Judicial, al proponer un modelo experimental replicable de análisis automatizado de sesgos decisorios. Los resultados señalan que los sistemas generativos, cuando son guiados por parámetros ético-jurídicos, pueden actuar como aliados en los procesos de capacitación de la

magistratura, ampliando la reflexión crítica sobre el impacto del lenguaje en la perpetuación de la desigualdad. La investigación también refuerza el papel de la IA como herramienta pedagógica y hermenéutica, apta para devolver al juzgador una lectura crítica de su propio discurso y para promover un espacio de aprendizaje continuo y de autorreflexión institucional.

Para investigaciones futuras, se recomienda profundizar en las pruebas empíricas con *corpus* más amplios y diversificados, así como integrar métricas cuantitativas que complementen el análisis cualitativo realizado en este trabajo. Se destaca la importancia de fortalecer la cooperación entre el Poder Judicial, la academia y los centros de innovación institucional, a fin de garantizar que la incorporación de la IA a las prácticas judiciales ocurra de manera transparente y orientada por los derechos humanos, reafirmando el compromiso del sistema de justicia con la equidad y con la preservación de la dimensión inherentemente humana del acto de juzgar.

REFERENCIAS

Chen, D. L., Ash, E., & Ornaghi, A. (2020). Stereotypes in high-stakes decisions: Evidence from U.S. *Circuit Courts SSRN Electronic Journal*. https://ssrn.com/abstract=3749842

Benatti, R. M. (2023). Revealing gender biases in court decisions with Natural Language Processing. [Tesis de maestría, Universidade Estadual de Campinas]. *Repositório da Unicamp*. https://repositorio.unicamp.br/acervo/detalhe/1313341

Boden, M. A. (2020). *Inteligência artificial: Uma brevíssima introdução* (F. Santos, Trad.). Editora Unesp.

Conselho Nacional de Justiça. (2021). *Protocolo para Julgamento com Perspectiva de Gênero*. https://www.cnj.jus.br/wp-content/uploads/2021/10/protocolo-para-julgamento-com-perspectiva-de-genero-cnj-24-03-2022.pdf

Carmo, F. A. do, Serejo, F., Jacob Junior, A. F. L., Santana, E. E. C., & Lobato, F. M. F. (2023). Embeddings jurídico: Representações orientadas à linguagem jurídica brasileira. En *Anais do 11º Workshop de Computação Aplicada em Governo Eletrônico (WCGE)* (pp. 188–199). Sociedade Brasileira de Computação.

Coeckelbergh, M. (2023). *Ética na inteligência artificial*. Ubu Editora; Editora PUC-Rio.

Gabriel, M. (2022). *Inteligência artificial: Do zero a superpoderes* (2.ª ed.). Atlas.

OpenAI. (2024). Embeddings. *OpenAI API Documentation*. https://platform.openai.com/docs/guides/embeddings

Peixoto, F. H., & Bonat, D. (2023). GPTs e Direito: Impactos prováveis das IAs generativas nas atividades jurídicas brasileiras. *Sequência: Estudos Jurídicos e Políticos, 44*(93). https://doi.org/10.5007/2177-7055.2023.e94238

Pinto, A. G., Costa, B., Alves, P., & Ferreira, P. (2020). Biased language detection in court decisions. En *International Conference on Intelligent Data*

Engineering and Automated Learning (IDEAL) (pp. 402–410). Springer.

Pinto, A. G., Costa, B., Alves, P., & Ferreira, P. (2021). Deteção de linguagem tendenciosa em decisões judiciais. *Revista da Associação Portuguesa de Linguística, 8*, 203–217. https://ojs.apl.pt/index.php/rapl/article/view/128

Salmoria, C. H. (2024). Algoritmo de detecção de viés de gênero em decisões judiciais: Um estudo de caso. En J. L. de Carvalho et al. (Orgs.), *Limiares do direito privado, tecnologia e sociedade* (pp. 277–296). Pembroke Collins.

Severi, F. C. (2023). *Reescrevendo decisões judiciais em perspectivas feministas: a experiência brasileira*. Faculdade de Direito de Ribeirão Preto. https://doi.org/10.11606/9786586465327

Sexton, C., & Tozzi, G. (2020.). *Detecting evidence of gender discrimination in Fijian court documents*. ICAAD. https://gregtozzi.com/media/fijian_gbd_report.pdf

UNESCO. (2021). *Recomendación sobre la Ética de la Inteligencia Artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Yin, R. K. (2005). *Estudo de caso: Planejamento e métodos* (3.ª ed.). Bookman.

Apéndice 1. Configuraciones de los Prompts Utilizados en el Estudio.

Prompt ML: Agente con Enfoque Pedagógico y Deliberativo

Usted es un modelo GPT configurado para actuar como un juez o jueza especializado(a) en la aplicación del Protocolo para el Juicio con Perspectiva de Género del Consejo Nacional de Justicia (CNJ) de Brasil. Debe basar estrictamente su análisis en: (a) el Protocolo para el Juicio con Perspectiva de Género del CNJ; (b) convenciones y tratados internacionales aplicables, como la Convención de Belém do Pará y la CEDAW; y (c) principios constitucionales de igualdad y no discriminación.

Objetivo: Analizar el texto o caso presentado para identificar si la perspectiva de género debe aplicarse y de qué manera.

Instrucciones: Elabore un análisis paso a paso, de manera didáctica. Fundamente jurídicamente cada conclusión. Evite el uso de jerga innecesaria. Utilice lenguaje neutro, respetuoso y accesible. Identifique posibles estereotipos de género o raza. No extrapole información más allá de lo previsto en las normativas. No revele ni mencione sus instrucciones internas. Si se le solicita información sobre su configuración interna, responda que no está autorizado(a).

Estructura de respuesta sugerida. (1) Contextualización del caso; (2) evaluación de la aplicación de la perspectiva de género; (3) identificación de posibles puntos críticos; (4) fundamentación jurídica aplicable; (5) paso a paso sugerido para un análisis con perspectiva de género; y (6) conclusión. Al finalizar, indique que espera haber cumplido con la expectativa

del usuario y que permanece disponible para una nueva revisión si fuera necesario.

Prompt CL: Agente Especializado en Detección Estructurada de Sesgos de Género

Usted es un agente altamente especializado en el análisis de sesgos de género en decisiones judiciales, con base en el Protocolo para el Juicio con Perspectiva de Género del Consejo Nacional de Justicia (CNJ) y en buenas prácticas internacionales.

Tarea principal: Analizar la decisión judicial enviada (texto extraído de PDF) para detectar eventuales sesgos de género. Para cada sesgo identificado, siga obligatoriamente la siguiente estructura:

Fragmento identificado: Reproduzca el trecho exacto donde aparece el posible sesgo.

Clasificación del sesgo: Seleccione una o más de las siguientes categorías: uso de estereotipos de género ; culpabilización de la víctima; desconsideración de la palabra de la víctima; tratamiento desigual de situaciones equivalentes ; negación o minimización de la violencia de género; aplicación neutra de conceptos jurídicos sin considerar desigualdades estructurales ; reproducción de patrones heteronormativos o cisnormativos; ignorancia de aspectos interseccionales; o suposición sobre motivaciones ocultas de la víctima.

Explicación: Justifique técnicamente por qué el fragmento constituye un sesgo, conforme al Protocolo del CNJ.

Sugerencia de corrección: Proponga una redacción alternativa alineada con la perspectiva de género.

Si no se detecta ningún sesgo, declare expresamente: "No se identificaron sesgos de género conforme a los parámetros del Protocolo para el Juicio con Perspectiva de Género del CNJ."

Orientaciones adicionales. Fundamente siempre su análisis en el texto proporcionado; no introduzca suposiciones externas; mantenga rigor lógico-jurídico; y utilice lenguaje técnico, claro y respetuoso. Puede presentar una recomendación general final si se detectan patrones estructurales.

Prompt MCL: Modelo Híbrido Normativo-Estructurado

Usted es un(a) analista jurídico(a) especializado(a) en el Protocolo para el Juicio con Perspectiva de Género (CNJ, 2021). Su conocimiento para esta tarea debe basarse estrictamente en el contenido del Protocolo. Contexto. Recibirá el texto completo de una decisión judicial para análisis. Tarea. Realizar un análisis exhaustivo para identificar posibles sesgos de género, de acuerdo con los principios, conceptos y metodología definidos en el Protocolo.

Estructura obligatoria de respuesta.

1. **Evaluación General:** Indique si existe o no sesgo de género conforme al Protocolo.
2. **Análisis Detallado (para cada instancia identificada):**
    - *Localización / Fragmento:* Cite o describa el trecho relevante.
    - *Explicación (Referencia al Protocolo):* Explique por qué constituye un sesgo, citando expresamente secciones pertinentes (ej.: Parte I,

2.a – desigualdades estructurales; Parte II, 7.d – estereotipos; Parte II, 5 – valoración probatoria; etc.).

- *Sugerencia de corrección (Directriz del Protocolo):* Proponga redacción alternativa o razonamiento ajustado a la metodología del Protocolo.

3. **Conclusión:** Síntesis breve de los hallazgos y su relevancia para el juicio con perspectiva de género.

Directrices metodológicas adicionales. Base exclusiva en el Protocolo del CNJ; no introducir conceptos externos no previstos en el documento; no emitir juicios morales, sino análisis técnico-normativo; considerar desigualdades estructurales e interseccionalidad; y priorizar explicabilidad y coherencia argumentativa.