Poder, Vigilancia y Control en la Era Algorítmica: La Inteligencia Artificial Estatal Frente a las Garantías de Igualdad y dignidad en el Estado Social de Derecho

Luis Enrique Echeverria Ahumada

Trabajo de Grado para Optar el Título de Abogado

Director
René Álvarez Orozco
Magister en Historia

Universidad Industrial De Santander
Facultad de Ciencias Humanas
Escuela de Derecho y Ciencia Política
Bucaramanga
2026

Tabla de Contenido

Lista de Tablas

Resumen

Esta investigación analiza cómo la incorporación de sistemas de inteligencia artificial en dispositivos estatales de poder, vigilancia y control reconfigura las garantías de igualdad material, no discriminación y dignidad humana en el Estado Social de Derecho. El problema de investigación se sitúa en la incertidumbre sobre las condiciones bajo las cuales estas tecnologías, presentadas bajo promesas de eficiencia y neutralidad, reproducen o mitigan desigualdades históricas mediante el uso de datos sesgados, modelos opacos, arreglos de gobernanza insuficientes y controles humanos débiles. En consecuencia, el objetivo principal consiste en determinar, por medio de qué mecanismos técnico-institucionales la IA estatal preserva, erosiona o transforma dichas garantías. Metodológicamente, la tesis adopta un enfoque jurídico-hermenéutico, cualitativo y documental, articulado con un estudio comparado de casos. A partir de matrices de caracterización, comparación y evaluación garantista, examina cinco configuraciones sociotécnicas concretas: COMPAS, SyRI, AFR Locate, PRiSMA y Fiscal Watson. El desarrollo del trabajo reconstruye, primero, el marco conceptual sobre IA, poder, vigilancia y control; luego caracteriza la operación técnica e institucional de los casos; después identifica umbrales mínimos de validez constitucional; y, finalmente, evalúa cada caso según su nivel de compatibilidad garantista. La investigación concluye que la IA estatal no afecta las garantías de forma abstracta ni uniforme, sino según su configuración concreta, y que, en los casos analizados, las promesas de neutralidad tienden a encubrir mecanismos de reproducción de desigualdad. Ninguno de los sistemas estudiados satisface plenamente el conjunto de garantías evaluadas, por lo que su validez constitucional exige condiciones robustas de transparencia, trazabilidad, auditoría, evaluación de impacto y control humano significativo.

Palabras clave: inteligencia artificial, Estado social de Derecho, igualdad material, no discriminación, dignidad humana, control humano significativo.

Abstract

This thesis analyzes how the incorporation of artificial intelligence systems into state apparatuses of power, surveillance, and control reconfigures the guarantees of substantive equality, non-discrimination, and human dignity within the Social Rule of Law. The research problem lies in the uncertainty surrounding the conditions under which these technologies, presented through promises of efficiency and neutrality, reproduce or mitigate historical inequalities through biased data, opaque models, insufficient governance arrangements, and weak human oversight. Accordingly, the main objective is to determine, through which techno-institutional mechanisms state AI preserves, erodes, or transforms those guarantees. Methodologically, the thesis adopts a legal-hermeneutic, qualitative, and documentary approach, articulated through a comparative case study. Based on characterization, comparison, and rights-based evaluation matrices, it examines five concrete sociotechnical configurations: COMPAS, SyRI, AFR Locate, PRiSMA, and Fiscal Watson. The study first reconstructs the conceptual framework on AI, power, surveillance, and control; then characterizes the technical and institutional operation of the cases; subsequently identifies minimum thresholds of constitutional validity; and finally evaluates each case according to its level of rights-based compatibility. The research concludes that state AI does not affect guarantees in an abstract or uniform manner, but rather according to its specific configuration, and that, in the cases analyzed, promises of neutrality tend to conceal mechanisms that reproduce inequality. None of the systems studied fully satisfies the set of guarantees assessed; therefore, their constitutional validity requires robust conditions of transparency, traceability, auditing, impact assessment, and meaningful human control.

Keywords: artificial intelligence, social Rule of Law, substantive equality, non-discrimination, human dignity, meaningful human control

Introducción

El auge de la Inteligencia Artificial (IA)[1] se ha consolidado como uno de los fenómenos más transformadores e intrigantes de nuestro tiempo. Su influencia se extiende a múltiples ámbitos, incluyendo la economía, la cultura, la política y la gestión estatal. La IA ha dejado de ser una mera herramienta para convertirse en un componente estructural de la vida contemporánea, hasta el punto de ser presentada en diversos escenarios como un recurso capaz de reemplazar al ser humano en la toma de decisiones.

Esta presencia se advierte en contextos como el colombiano, donde la IA se incorpora paulatinamente en la vida cotidiana, en el terreno profesional para la gestión de datos y la automatización de procesos, y en el diseño de políticas públicas para mejorar la eficiencia administrativa. De manera particular, en el ámbito jurídico, la IA se ha incorporado como apoyo en la toma de decisiones, desde la gestión documental hasta sistemas orientados a la predicción de sentencias. No obstante, estos usos suscitan interrogantes de mayor calado: la tecnología, lejos de ser neutra, tiende a reproducir y amplificar los sesgos inscritos en los datos con los que es entrenada, reflejo de procesos históricos de discriminación y exclusión.

Estos interrogantes se agudizan si se considera que la expansión vertiginosa de la Inteligencia Artificial se articula con algoritmos opacos, a menudo diseñados bajo criterios éticos, morales y jurídicos cuestionables, y atravesados por sesgos que reconfiguran las dinámicas sociales e individuales. Todas estas transformaciones plantean retos inéditos al

[1] En lo sucesivo, la sigla IA se empleará para referirse a Inteligencia Artificial.

Derecho y al Estado Social de Derecho (ESD)[2], sus garantías fundamentales y la redefinición de los procesos de decisión confiados a la administración pública y judicial.

Estos retos no son meramente abstractos, sino que se materializan y generan consecuencias evidentes en nuestra realidad. En este sentido, ejemplos paradigmáticos como SyRI en Países Bajos —anulado por discriminar a grupos vulnerables—, el reconocimiento facial en vivo en el Reino Unido —cuestionado por sus altos índices de falsos positivos y su sesgo desproporcionado contra minorías raciales—, COMPAS en Estados Unidos —criticado por atribuir mayor riesgo de reincidencia a personas negras frente a personas blancas en situaciones similares— y, en Colombia, PRISMA —suspendido por falta de garantías— y el denominado Fiscal Watson muestran que aquellas herramientas presentadas como soluciones técnicas eficientes pueden, en realidad, operar sobre bases de datos sesgados, mediante modelos opacos y con escasos mecanismos de control democrático y en consecuencia, funcionar como dispositivos de exclusión y dominación.

De este modo, la presente investigación se centra en analizar cómo la incorporación de sistemas de IA en dispositivos estatales de poder, vigilancia y control reconfigura las garantías del Estado Social de Derecho. La delimitación temática se enfoca en los mecanismos (diseño, datos, gobernanza y control humano significativo) mediante los cuales la IA reproduce o mitiga desigualdades y, con ello, incide en los principios de igualdad material, no discriminación y dignidad humana.

En este marco, la expansión de sistemas de IA en la gestión pública y de la seguridad se inserta en tramas de poder, vigilancia y control apoyadas en infraestructuras de datos con

[2] En lo sucesivo, la sigla ESD se empleará para referirse Al Estado Social de Derecho.

huellas históricas de desigualdad. Esto adquiere una especial relevancia en contextos como el colombiano, donde el poder punitivo ha estado históricamente ligado al control social de los sectores más vulnerables, de modo que la introducción acrítica de modelos algorítmicos que, sin debate público suficiente ni controles materiales sobre sus efectos, corre el riesgo de profundizar y consolidar las desigualdades ya existentes.

En este contexto, la investigación parte de analizar si se está configurando una forma de control y castigo mediada por algoritmos que, bajo el lenguaje de la neutralidad técnica, reproduce prácticas de vigilancia y castigo selectivos. Desde esta perspectiva, estos mismos sistemas —SyRI, el reconocimiento facial en el Reino Unido, COMPAS, PRiSMA y el Fiscal Watson— se abordarán como estudios de caso que permiten examinar cómo estos sistemas se insertan en las tramas de poder, vigilancia y control social. De allí que, en términos de relevancia social y práctica, este trabajo resulte oportuno al proponer un marco integrado para evaluar si las promesas de eficiencia de la IA encubren y refuerzan sesgos estructurales o si, por el contrario, pueden ser reorientadas mediante salvaguardas robustas que eleven el estándar de protección de derechos.

En consecuencia, el objetivo general de esta tesis es analizar, mediante un estudio de casos nacionales e internacionales**,** en qué medida y por medio de qué mecanismos (diseño, datos, gobernanza y control humano significativo) la incorporación de sistemas de IA en dispositivos estatales de poder, vigilancia y control reproduce o mitiga desigualdades, y cómo reconfigura las garantías del Estado Social de Derecho (igualdad material, no discriminación y dignidad humana). Para lograr esto, el estudio reconstruirá el marco conceptual, caracterizará las configuraciones técnico-institucionales de los casos seleccionados, comparará los hallazgos para reconstruir criterios de validez constitucional, y

evaluará la incidencia de dichas configuraciones en las garantías del Estado Social de Derecho.

1. Planteamiento y Formulación del Problema

El auge de la Inteligencia Artificial (IA) viene constituyendo uno de los fenómenos más intrigantes y transformadores de nuestros tiempos. Su presencia se extiende a múltiples ámbitos de la vida contemporánea, desde la economía y la cultura hasta la política y la gestión estatal Su influencia lejos de permanecer en el estricto ámbito profesional ha traspasado a la vida social e individual, alcanzando incluso la esfera del entendimiento personal. Allí, a través de algoritmos opacos, —con frecuencia diseñados bajo criterios éticos, morales y jurídicos cuestionables, y atravesados por sesgos que favorecen intereses de grandes corporaciones o Estados—, reconfigura dinámicas de interacción, comunicación e incluso la construcción de identidad (Santana, 2025).

Sin embargo, más allá de la valoración que pueda hacerse sobre sus impactos favorables o desfavorables, la Inteligencia Artificial constituye ya un componente estructural de la vida contemporánea, no solo como herramienta, sino como un agente capaz de reemplazar al hombre en la toma de decisiones, cuya permanencia en el tiempo parece indiscutible, como si tratase de una predestinación contra la cual no se puede luchar (Harari, 2024).

Este carácter irreversible e inevitable no solo se advierte en los países con mayor desarrollo tecnológico, sino también en contextos como el colombiano, donde la Inteligencia Artificial comienza a incorporarse de manera paulatina en diversos ámbitos. En la vida

cotidiana, su presencia se refleja en aplicaciones de uso masivo que median la comunicación, el consumo y el acceso a la información (DNP, 2023; DANE, 2025). En el terreno profesional, empieza a ser utilizada como herramienta de apoyo en la toma de decisiones, la gestión de datos y la automatización de procesos (DNP, 2023; Herrera Giraldo et al., 2024). Incluso en el diseño de políticas públicas, se observa un interés creciente por integrar estas tecnologías para mejorar la eficiencia administrativa y la prestación de servicios (PNUD, 2024; CONPES 4144, 2025).

Todas estas transformaciones no resultan ajenas al Derecho, ciencia en la cual la inteligencia artificial plantea retos inéditos en materia de regulación, garantías fundamentales y redefinición de los procesos de decisión judicial, todos los cuales atañen a la gestión de los asuntos comunes de la sociedad, tradicionalmente confiados a la administración pública (Sarrión, 2023). En el ámbito jurídico, la inteligencia artificial se ha ido incorporando como herramienta de apoyo en la toma de decisiones judiciales, con aplicaciones que van desde la gestión documental y el tratamiento de grandes volúmenes de información y datos, hasta sistemas orientados a la predicción de sentencias (Asís, 2023).

Sin embargo, como señala Rafael de Asís (2023), estos usos, en apariencia neutros, suscitan interrogantes de mayor calado: lejos de garantizar imparcialidad, la Inteligencia Artificial tiende a reproducir y amplificar los sesgos inscritos en los datos con los que es entrenada, sesgos que a su vez son reflejo de procesos históricos de discriminación, exclusión y desigualdad social. En ese mismo sentido, referido autor advierte que "*No hay que pasar por alto que las aplicaciones de inteligencia artificial trabajan en función de los datos y la información que se les proporcionan y que, si no se corrigen, reproducen lógicas discriminatorias*" (Asís, 2023, p.26).

Así, lo que se presenta como una herramienta objetiva y técnica puede terminar consolidando estigmas y jerarquías sociales preexistentes, bajo el ropaje de una supuesta neutralidad algorítmica (Innerarity, 2025).

De este modo, principios esenciales del Estado Social de Derecho como la igualdad material, la no discriminación y la dignidad humana (Asís, 2023), así como la axiología del castigo (Hernández, 2025), se ven amenazados por un modelo tecnológico que en lugar de corregir desigualdades estructurales corre el riesgo de perpetuarlas (Presno, 2025). En esa línea, Innerarity advierte que "*Los efectos discriminatorios de la aplicación de la inteligencia artificial no sólo discriminan de hecho, sino que generan un efecto de retroalimentación: los prejuicios se asientan por la correspondiente confirmación de unas máquinas supuestamente neutrales"* (2025, p. 334).

En este sentido, la IA debe comprenderse no solo como un recurso técnico, sino también como un nuevo dispositivo de poder que se articula con las lógicas estatales de disciplina, vigilancia y control, proyectando sus efectos tanto sobre los cuerpos como sobre las subjetividades, y configurándose como una prolongación sofisticada de los mecanismos de dominación social ya existentes (Hernández, 2025).

Esta tensión no es exclusiva del caso colombiano, pues a nivel global los gobiernos han comenzado a implementar sistemas algorítmicos para optimizar procesos administrativos, vigilar el cumplimiento de normas y gestionar recursos, bajo la promesa de una mayor eficiencia, rapidez y neutralidad en la toma de decisiones (Rivero, 2023; Innerarity, 2025). No obstante, este despliegue tampoco está exento de riesgos: el uso de herramientas tecnológicas para controlar a la población plantea interrogantes sobre poder,

transparencia y legitimidad, en tanto la aparente objetividad técnica puede encubrir relaciones de exclusión y dominación (Innerarity, 2025).

La narrativa de la eficiencia técnica convive entonces con críticas que advierten sobre la opacidad de los algoritmos (Innerarity, 2025), la dificultad de impugnarlos (Hernández, 2025) y su potencial para reproducir desigualdades estructurales (Innerarity, 2025). Ejemplos de ello se han evidenciado en distintos países: el sistema SyRI en Países Bajos, anulado por discriminar a grupos vulnerables[3] (Bekkum & Zuiderveen, 2021) ; el uso de reconocimiento facial en vivo en Reino Unido, declarado ilegal[4] (R (Bridges) v Chief Constable of South

[3] SyRI fue un sistema neerlandés de detección de fraude en la asistencia social que enlazaba grandes volúmenes de datos públicos para generar notificaciones de riesgo. "*Una notificación de riesgo implica que se considera que una persona física o jurídica merece ser investigada por posible fraude, uso ilícito o incumplimiento de la legislación. Este método permite un uso más eficaz y eficiente del instrumento de control*" (van Bekkum & Zuiderveen, 2021, p. 325). Su diseño resultaba estructuralmente opaco, es decir, con variables y lógica no accesibles para terceros, ausencia de deber de información a los afectados y con límites difusos respecto de los principios de finalidad y minimización de datos. Bajo la promesa de eficiencia, el esquema corría el riesgo de "*automatizar la sospecha*", la cual fue denunciado por la campaña "Bij Voorbaat Verdacht", cuya traducción más fiel es "Sospechoso de antemano". Esta expresión señala que, al cruzar masivamente bases de datos y asignar puntuaciones de riesgo, ciertos colectivos podían ser tratados como sospechosos desde el inicio, antes de cualquier evidencia individual, lo que desplazaba garantías esenciales de transparencia y control ciudadano. En la práctica, SyRI se aplicó sobre todo en barrios empobrecidos, facilitando estigmatización territorial y riesgos de discriminación indirecta (sesgos socioeconómicos y migratorios). En febrero de 2020, el Tribunal de La Haya declaró ilegal su base jurídica por vulnerar el art. 8 CEDH en el sentido en que faltó un "*justo equilibrio*" entre interés público e injerencia en la vida privada, con transparencia insuficiente, uso desproporcionado de datos y evaluaciones de impacto inadecuadas. El gobierno no apeló y SyRI dejó de usarse. La lección de este caso es que sin explicabilidad y proporcionalidad, la gobernanza algorítmica pública erosiona derechos y reproduce desigualdades (van Bekkum & Zuiderveen, 2021).

[4] En el Reino Unido, la policía de Gales del Sur llevo a cabo el uso de cámaras de reconocimiento facial en vivo (AFR Locate), que capturaban rostros en tiempo real, generaban plantillas biométricas y las comparaban con "watchlists" de personas buscadas. Según su finalidad declarada, la herramienta buscaba eficiencia en seguridad pública en la medida en que captaban la imagen de los transeúntes, generaban una plantilla biométrica y la comparaban con las de la lista de los individuos buscados: si coincidía, se activaba una alerta. Sin embargo, su funcionamiento implicaba la captación indiscriminada de transeúntes no investigados, criterios opacos para definir las listas de búsqueda, lugares de despliegue, y riesgos técnicos (falsos positivos, sesgos por iluminación, ángulos o datos de entrenamiento). Críticamente estos rasgos afectan privacidad (captura masiva no consentida) e

Wales Police & Others, 2020); en Colombia, el cuestionado modelo PRISMA de la Fiscalía, orientado a predecir reincidencia y funcionalmente emparentado con sistemas de evaluación de riesgo como COMPAS[5] en Estados Unidos, que fue suspendido por falta de garantías[6]

igualdad (posibles errores y sesgos desigualmente distribuidos), al tiempo que dificultaba la rendición de cuentas sobre por qué alguien fue señalado (R (Bridges) v Chief Constable of South Wales Police & Others, 2020). La Corte de Apelación declaró ilegal este uso por no estar "conforme a derecho" por su marco normativo impreciso y discrecionalidad excesiva en quién incluir y dónde debía ser desplegado el sistema, por deficiencias en la evaluación de impacto en protección de datos y por incumplir el deber de igualdad del sector público. En síntesis, el caso sienta una lección replicable: las tecnologías de vigilancia algorítmica requieren base legal clara y específica, criterios verificables de necesidad y proporcionalidad, transparencia sobre datos y modelos, evaluaciones de impacto robustas y vías reales de recurso; sin estos estándares, su adopción tiende a normalizar la vigilancia y reproducir desigualdades.

[5] COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) es un sistema de evaluación de riesgo utilizado en distintas jurisdicciones de Estados Unidos para estimar la probabilidad de reincidencia de personas procesadas o condenadas. Desarrollado por la empresa privada Equivant, su objetivo declarado es apoyar decisiones sobre libertad bajo fianza, imposición de penas y libertad condicional, asignando puntajes de riesgo (bajo, medio o alto) a partir de cuestionarios y datos penales y sociodemográficos. Bajo la promesa de objetividad y eficiencia, COMPAS encarna la misma lógica actuarial que inspira herramientas como PRISMA: traducir trayectorias vitales y conflictos sociales en perfiles estadísticos que orientan la intervención penal. Sin embargo, investigaciones empíricas, han mostrado que el sistema presenta sesgos sistemáticos: tiende a sobrestimar el riesgo de reincidencia en personas negras y a subestimarlo en personas blancas, pese a tener tasas reales de reincidencia similares, lo que pone en cuestión su neutralidad y agrava desigualdades raciales preexistentes. A ello se suma su carácter de "caja negra": el algoritmo es de propiedad privada sujeta al secreto comercial, la ponderación de variables no es transparente y resulta extremadamente difícil impugnar o auditar sus resultados, lo que compromete garantías como el debido proceso, la presunción de inocencia y el derecho a la defensa. En consecuencia, COMPAS se ha convertido en un ejemplo paradigmático de "penalidad algorítmica": una forma de gobierno punitivo mediado por modelos opacos que, lejos de corregir la selectividad del sistema penal, puede reforzarla y legitimarla bajo el lenguaje tecnocrático de la predicción y el riesgo (Ibáñez, 2023; Cancio, 2023).

[6] PRiSMA (Perfil de Riesgo de Reincidencia para la Solicitud de Medidas de Aseguramiento) fue una herramienta automatizada desarrollada por la Fiscalía General de la Nación de Colombia para estimar la probabilidad de que una persona vuelva a delinquir. Su propósito oficial es servir como soporte técnico en audiencias de medida de aseguramiento, aportando un "perfil de riesgo" con base en datos de la Policía, Fiscalía e INPEC. El uso se justificó bajo la lógica de eficiencia y optimización de recursos judiciales, aunque introduce tensiones sobre cómo equilibrar ese fin con los derechos fundamentales de los investigados. El funcionamiento de PRiSMA implicó alimentar un modelo de aprendizaje automático con múltiples variables personales, antecedentes y eventos delictivos pasados, para generar una puntuación de riesgo. Pero su operación fue opaca en cuanto a los criterios de peso que asigna el algoritmo, así como los umbrales que derivan en decisiones. En la práctica, esta opacidad puede reforzar sesgos estructurales, estigmatizar personas de barrios marginales y dificultar

(Universidad de los Andes, 2024) y el programa WATSON[7] (Palacios et al., 2024), aún vigente, pero con numerosas críticas sobre su uso y funcionalidad.

Estos antecedentes, permiten advertir que los riesgos asociados a la IA no se limitan al ámbito administrativo o de vigilancia ciudadana, sino que adquieren una especial gravedad cuando se trasladan al campo del poder del control social (Innerarity, 2025). En efecto, la política criminal, históricamente ligada al control social de los sectores más vulnerables (Pawlik, 2022), se enfrenta al desafío de integrar nuevas tecnologías que prometen reducir la reincidencia y aumentar la eficacia del sistema (Rivero, 2023).

Sin embargo, el riesgo es que la inteligencia artificial termine reforzando las *funciones encubiertas de la pena* —noción desarrollada por Leal (2021) para dar cuenta de los fines no declarados del castigo, esto es, la estigmatización y exclusión de determinados

la impugnación eficaz de los resultados. Según estudios críticos, la herramienta pudo vulnerar derechos como la presunción de inocencia, el debido proceso y la igualdad material. Aunque no es claro el estado actual de PRiSMA, su empleo sin salvaguardas fuertes representa un precedente peligroso: sin transparencia y proporcionalidad, la predicción algorítmica penal puede desplazar imperceptiblemente las decisiones humanas y consolidar desequilibrios de poder.

[7] Fiscal Watson es el nombre dado por la Fiscalía General de la Nación a la adopción de IBM Watson (Explorer y luego Discovery) para indexar y analizar información del SPOA y otras fuentes, con fines de asociación de casos y apoyo a la investigación. Documentos contractuales le atribuyen funcionalidades como sugerir modus operandi, actuaciones, tipo de delito y vocación de éxito, además de apoyo en la asignación de casos. Aunque la entidad lo presenta como un motor de búsqueda analítica para agilizar el trabajo, la información pública sobre su gobernanza es escasa: no hay análisis de impacto en derechos publicados, ni métricas de uso y efectividad; y persiste opacidad sobre variables, modelos y umbrales aplicados. En la práctica, la Fiscalía anunció su uso desde 2018 y exhibió "casos de éxito" por asociaciones rápidas de denuncias; sin embargo, medios y organizaciones han planteado riesgos de soberanía de la información, dependencia de proveedor y transparencia insuficiente. La cadena de renovaciones por contratación directa con IBM y la centralidad de su infraestructura refuerzan temores de "caja negra" y dificultades para escrutinio, auditoría externa e impugnación de resultados que orienten decisiones en el marco del proceso penal, en consecuencia está latente la vulneración a derechos fundamentales tales como derecho al debido proceso y a un juicio justo, privacidad y protección de datos, el derecho a recursos efectivos, a la prohibición de discriminación, entre otros (Palacios et al., 2024).

grupos sociales—, bajo el ropaje de una neutralidad técnica (Innerarity, 2025). El caso colombiano ilustra con claridad este dilema: en un sistema penitenciario ya en crisis, marcado por el hacinamiento estructural y por la distancia entre los fines legítimos de la pena y aquellos que se materializan en la práctica —los denominados *fines encubiertos*—, la introducción de modelos algorítmicos sin una regulación suficiente puede profundizar las desigualdades en lugar de corregirlas (Santana, 2025).

Por lo tanto, la expansión de sistemas de IA hacia la política criminal y la administración pública en Colombia y el mundo, ocurre entre datos sesgados, modelos opacos y marcos de control insuficientes. No está claro en qué condiciones tales sistemas reproducen (por sesgos, opacidad y discrecionalidad) o mitigan (mediante transparencia, control humano significativo y evaluación de impactos) patrones históricos de exclusión y desigualdad[8] que caracterizan al ejercicio del poder punitivo y, en general, a las estructuras de control social contemporáneas. En su deriva más crítica, la incorporación acrítica de la IA podría consolidar nuevas formas de disciplinamiento y vigilancia que, lejos de democratizar la justicia o fortalecer las garantías del Estado Social de Derecho, profundicen las asimetrías y reproduzcan las lógicas de dominación que las tecnologías prometían superar.

[8] En términos analíticos, la incertidumbre planteada ("no está claro en qué condiciones reproducen o mitigan patrones históricos de exclusión y desigualdad") se abordará mediante un conjunto de preguntas guía, aplicables a cada estudio de caso, que permitirán reconstruir de manera comparable las configuraciones sociotécnicas y sus efectos. En particular, para cada sistema se indagará: (i) qué datos utiliza (fuentes, selección, tratamiento y posibles proxies sensibles); (ii) qué hace el modelo y cómo interviene en la práctica (puntúa, clasifica, perfila, prioriza o genera alertas, y con qué peso real en la decisión estatal); (iii) cuál es el grado de opacidad o explicabilidad relevante para el caso (qué se conoce sobre variables, lógica y límites); (iv) cómo se estructura su gobernanza institucional (base jurídica invocada, operador/es, controles, auditorías y trazabilidad); (v) si existe control humano significativo (capacidad real de apartarse del output o resultado, motivación y registro de la decisión final, y revisión); y (vi) qué impactos documentados o plausibles se proyectan sobre igualdad material, no discriminación y dignidad humana.

Con ello, la expansión de sistemas de IA en la gestión pública y de la seguridad se inserta en tramas de poder, vigilancia y control, apoyadas en infraestructuras de datos con huellas históricas de desigualdad, modelos opacos y mecanismos débiles de rendición de cuentas. En tales condiciones, su despliegue puede según su diseño, contexto y gobernanza, reproducir desigualdades y exclusiones o, bajo salvaguardas robustas, mitigarlas, con efectos que se proyectan tanto en los circuitos administrativos de control y vigilancia estatales como, de forma paradigmática, en el dispositivo punitivo y de encierro.

En el mismo sentido no está claro, en qué configuraciones institucionales y técnico-metodológicas llevan a la IA amplificar asimetrías (por opacidad, sesgos y discrecionalidad) o corregir injusticias (mediante transparencia, control humano significativo y evaluación de impactos), ni cómo ello reconfigura las garantías del Estado Social de Derecho (igualdad material, no discriminación y dignidad humana). De ahí la necesidad de examinar, caso por caso, las condiciones bajo las cuales estas tecnologías se alinean con un horizonte garantista o, por el contrario, profundizan lógicas de dominación.

A efectos comparativos, el estudio no contrapondrá jurisdicciones, sino configuraciones sociotécnicas: combinaciones de datos, modelo, regla de uso en el proceso decisorio y gobernanza institucional. La comparación buscará identificar qué rasgos (por ejemplo, opacidad estructural, uso de proxies sensibles, ausencia de control humano verificable o carencia de recursos efectivos) aparecen recurrentemente cuando se erosionan garantías, y qué rasgos (evaluaciones de impacto robustas, trazabilidad, transparencia operativa, auditorías y control humano significativo) correlacionan con escenarios de mitigación y preservación.

En consecuencia, a partir de estos hallazgos empíricos sobre mecanismos de reproducción o mitigación de desigualdades en criterios jurídicos verificables, el trabajo se orienta a un análisis crítico del modo en que la IA incide sobre la igualdad material, la no discriminación y la dignidad humana. La contribución del estudio no radica en prescribir soluciones ni diseñar políticas, sino en reconstruir estándares constitucionales de control, identificar umbrales de validez (igualdad material, no discriminación y dignidad humana) y describir con precisión las configuraciones técnico-institucionales bajo las cuales tales estándares se ven preservados, erosionados o transformados.

De forma tal que la pregunta que se plantea es: **¿Cómo reconfigura la incorporación de sistemas de IA en dispositivos estatales de poder, vigilancia y control las garantías del Estado Social de Derecho —igualdad material, no discriminación y dignidad humana—, a partir de un análisis de casos nacionales e internacionales?**

Dicho problema de investigación queda entonces delimitado como la necesidad de mapear y contrastar caso por caso las formas en que la incorporación de sistemas de IA en dispositivos estatales de poder, vigilancia y control reconfigura las garantías propias del Estado Social de Derecho. El resultado esperado es un examen argumentado de mecanismos y efectos, junto con un conjunto de criterios de lectura jurídica que permitan valorar la conformidad (o disconformidad) de dichas prácticas con los preceptos constitucionales, sin derivar en recomendaciones regulatorias o de política pública.

2. Justificación

La investigación se justifica porque aborda con un diseño realista, un problema actual y especialmente relevante que se encuentra una zona crítica de incertidumbre jurídica: las condiciones bajo las cuales la incorporación de sistemas de Inteligencia Artificial en dispositivos estatales de poder, vigilancia y control, reproduce o mitiga desigualdades y, correlativamente, cómo ello reconfigura las garantías propias del Estado Social de Derecho de igualdad material, no discriminación y dignidad humana.

En términos de conveniencia, este trabajo es oportuno porque ofrece un marco integrado, sustentado en evidencia de casos y en criterios jurídicos verificables, para evaluar de manera sistemática si las promesas de eficiencia y neutralidad técnica de la IA encubren sesgos estructurales o, por el contrario, pueden orientarse mediante salvaguardas robustas que eleven el estándar de protección de derechos en el Estado Social de Derecho.

Además, la investigación desarrolla un examen argumentado de mecanismos y efectos, y propone criterios de lectura jurídica para valorar la conformidad o disconformidad de prácticas estatales de IA con los preceptos constitucionales, aportando claridad allí donde predominan decisiones tecnológicas opacas con efectos sobre derechos fundamentales. En esa medida, sirve a la academia y constituye una crítica rigurosa del ejercicio del poder público, especialmente en sus dispositivos de vigilancia, control y castigo, al traducir un problema complejo en pautas analíticas y criterios contrastables, sin convertirse en un recetario de política pública.

En términos de relevancia social, los resultados pueden beneficiar de manera diferenciada a personas y grupos históricamente vulnerados, al orientar la prevención de

formas estructurales de discriminación que la adopción de sistemas de IA puede agravar si no median salvaguardas adecuadas. También aportan a la comunidad académica y de investigación, al ofrecer un marco analítico y categorías operativas para estudiar críticamente la relación entre diseño, datos, gobernanza y garantías del Estado Social de Derecho. Finalmente, resultan útiles para operadores jurídicos, entidades estatales y órganos de control, porque brindan una lectura crítica de experiencias nacionales e internacionales que permite identificar patrones de riesgo, aprender de errores documentados y ajustar la gobernanza técnica y organizacional antes del despliegue o rediseño de sistemas, mediante umbrales de validez orientados a preservar efectivamente la igualdad y la dignidad.

Esta proyección social responde a las preguntas planteadas por de Sampieri & Mendoza (2018, p. 45) "*¿quiénes se beneficiarán?*" y *"¿de qué modo?*" como criterio para evaluar el valor potencial de un estudio. Desde la perspectiva de las implicaciones prácticas, aunque la tesis no prescribe políticas ni pretende desarrollar vías procesales para la implementación de la IA, su producto principal —un análisis crítico sobre la funcionalidad constitucional del uso de dispositivos de IA en entornos estatales— tiene usos inmediatos en la formación y el trabajo analítico de quienes diseñan, evalúan o supervisan el uso estatal de IA.

A sí mismo, ofrece herramientas analíticas para reconocer, a partir de experiencias previas nacionales e internacionales, cuándo ciertas configuraciones técnico-institucionales tienden a reproducir desigualdades o, por el contrario, a mitigarlas; delimita umbrales de validez y condiciones de legitimidad a la luz del Estado Social de Derecho, y orienta la toma de decisiones institucionales al anticipar riesgos y evitar repetir errores documentados.

En lugar de listas de cumplimiento, la matriz permite valorar, con criterios jurídicos verificables, si la priorización algorítmica redistribuye cargas y beneficios de modo compatible con la igualdad material y la dignidad, si la segmentación por perfiles deriva en formas de discriminación estructural y qué exigencias de control humano significativo se siguen de tales hallazgos. Así, el estudio responde a la cuestión de si "*¿ayudará a resolver alguno o varios problemas reales?*" (Sampieri & Mendoza, 2018, p. 45), en un sentido estrictamente jurídico-analítico: provee fundamentos para evaluar y corregir diseños y gobernanzas antes de su despliegue o rediseño, en consonancia con los objetivos de la investigación.

En su valor teórico, la investigación llena un vacío al desplazar el foco desde el rendimiento técnico hacia una reconstrucción constitucional de garantías en entornos tecnosociales: identifica cómo rasgos de diseño y datos pueden operar como proxies de categorías protegidas produciendo discriminación indirecta, cómo la clasificación automatizada puede devenir cosificante o estigmatizante aun con ganancias marginales de precisión, y qué estándares sustantivos delimitan la igualdad material y la dignidad cuando la asignación de recursos o riesgos se realiza por perfiles algorítmicos; esos hallazgos permiten derivar principios más amplios y formular hipótesis para investigaciones futuras, justo como exige el criterio de "valor teórico" de Sampieri & Mendoza (2018, p. 45).

La utilidad metodológica radica en que el diseño adoptado de estudio de casos en nacionales e internacionales es replicable y ofrece un esquema de análisis que puede aplicarse en otros contextos desde una perspectiva analítica y constitucional: rejillas de lectura para identificar relaciones entre diseño, datos, gobernanza y control humano significativo; tipologías de mecanismos de reproducción o mitigación de desigualdades;

matrices de correspondencia entre tales mecanismos y las garantías del Estado Social de Derecho (igualdad material, no discriminación y dignidad); y criterios de validez que permiten apreciar la legitimidad de configuraciones técnico-institucionales sin convertirlos en recetarios de cumplimiento.

Estos instrumentos no persiguen auditorías técnicas ni rutas procesales, sino definir con rigor conceptos, variables y relaciones y estudiar más adecuadamente el "objeto-población" de la investigación —las configuraciones de uso estatal de IA—, en sintonía con los lineamientos metodológicos de Sampieri & Mendoza (2018, p. 45).

En cuanto a la viabilidad y factibilidad, la tesis asume explícitamente —como recomienda Hernández-Sampieri— la evaluación de recursos, tiempo y alcance, circunscribiendo el estudio a un número acotado de casos (5) seleccionados por relevancia constitucional y diversidad de configuraciones (funcionalidad, arquitectura del modelo, gobernanza), dentro de un cronograma compatible con los requerimientos de un trabajo de grado; el acceso a fuentes resulta suficiente y razonable (decisiones judiciales, lineamientos administrativos, documentos públicos y doctrina), el diseño comparado con énfasis en el contexto colombiano maximiza el aprendizaje con costos económicos directos mínimos o nulos.

A la luz del planteamiento del problema y de los objetivos —analizar en qué medida y por medio de qué mecanismos la IA estatal reproduce o mitiga desigualdades y cómo reconfigura las garantías del ESDD, mediante un estudio de casos con énfasis en Colombia y contrastes internacionales— la investigación queda justificada por cumplir, de modo razonado y verificable, los criterios de evaluación del valor potencial de un estudio propuestos por Hernández-Sampieri (2018, p. 45) de conveniencia, relevancia social,

implicaciones prácticas, valor teórico y utilidad metodológica, por su originalidad comparada y por su viabilidad y delimitación realista para una tesis de grado.

3. Objetivo General y Objetivos Específicos

3.1 Objetivo general

Analizar mediante un estudio de casos con énfasis en Colombia y contrastes internacionales, en qué medida y por medio de qué mecanismos (diseño, datos, gobernanza y control humano significativo) la incorporación de sistemas de IA en dispositivos estatales de poder, vigilancia y control reproduce o mitiga desigualdades, y cómo reconfigura las garantías del Estado Social de Derecho (igualdad material, no discriminación y dignidad humana).

3.2 Objetivos específicos

1. Analizar la incorporación de sistemas de inteligencia artificial en dispositivos estatales de poder, vigilancia y control, con el fin de establecer las categorías analíticas que orientarán el desarrollo del problema central.

2. Identificar y caracterizar las configuraciones técnico-institucionales de los sistemas de IA en los casos seleccionados (fuentes de datos, diseño del modelo, esquemas de gobernanza y grado de control humano significativo), mapeando los principales mecanismos de afectación y de contención en relación con la (re)producción o mitigación de desigualdades en los dispositivos estatales analizados.

3. Examinar los hallazgos entre casos nacionales e internacionales para identificar las condiciones bajo las cuales las prácticas estatales con IA preservan, erosionan o transforman las garantías del Estado Social de Derecho y, a partir de ese examen, reconstruir los umbrales de validez constitucional de igualdad material, no discriminación y dignidad humana aplicables en el contexto analizado.

4. Evaluar cómo las configuraciones técnico-institucionales identificadas inciden en las garantías del Estado Social de Derecho —igualdad material, no discriminación y dignidad humana—, precisando los estándares aplicables a cada una y determinando, para cada caso, el nivel de cumplimiento y las desviaciones relevantes respecto de dichos estándares.

4. Capítulo I. La Inteligencia Artificial en los dispositivos Estatales de poder, vigilancia y control

4.1 Marco Teórico y Conceptual

La revisión de la literatura pertinente al planteamiento central y a los objetivos específicos ya propuestos revela un panorama complejo que abarca diversas tradiciones teóricas: derecho constitucional, sociología del control y del castigo y estudios sobre la tecnología y el poder. Por lo tanto, en atención a esta pluralidad, el presente apartado adopta una estructura fundamentada en diversos enfoques conceptuales para organizarse mediante el método de "vertebración" (Sampieri & Mendoza 2018, p. 89), es decir, de lo general a lo específico. Esta estructura permite abordar la complejidad y multiplicidad de detalles del castigo y la penalidad en la sociedad moderna.

4.1.1. Fundamentos conceptuales de la Inteligencia Artificial y su aplicación en el ámbito jurídico

En este apartado se presentan los principales fundamentos conceptuales de la Inteligencia Artificial (IA) y las formas en que ha sido empleada como modelo predictivo en el ámbito jurídico. Primero se aborda teóricamente el concepto de IA y, posteriormente, se examinan sus implicaciones como dispositivo de poder y control estatal.

4.1.1.1 Concepto, enfoques y tipologías de Inteligencia Artificial como tecnología de predicción.

No existe una definición universalmente aceptada del término "Inteligencia Artificial", menos aun cuando se comprende que es una expresión amplia y en constante evolución (Vázquez, 2025). Sin embargo, Kaplan (2017, citado en Vázquez Pita, 2025) señala que la mayoría de las definiciones convergen en referirse a programas o máquinas capaces de realizar conductas y/o resultados que se considerarían inteligentes si fueran efectuadas por humanos. De otro lado, operacionalmente la IA puede entenderse como sistemas basados en algoritmos[9] y autoaprendizaje que simulan el razonamiento humano, identifican problemas, procesan información y adoptan decisiones o efectúan acciones con cierto grado de autonomía (Sánchez, 2021).

Con todo ello, la IA se presenta como una disciplina de la informática, el análisis de datos, las estadísticas y la matemática (Hernández, 2025), cuyo objetivo es el desarrollo de

[9] El algoritmo es la secuencia de instrucciones que establece las acciones que la IA debe ejecutar para resolver un problema. Se compone de fórmulas matemáticas que, alimentadas por datos masivos, permiten a las computadoras realizar funciones avanzadas (Hernández, 2025).

máquinas y sistemas que puedan llevar a cabo tareas que normalmente requerirían intervención humana, como dar sentido al lenguaje o resolver problemas (Carballo, 2021, citado en Vázquez, 2025)

Sin embargo, la ola actual de avances en IA no se define tanto por aportar inteligencia humana, sino por ser una tecnología de la predicción, con la capacidad de aportar la información que falta a partir de los datos existentes (Agrawal, 2019, citado en Vázquez Pita, 2025). Estos sistemas, que funcionan con diferentes grados de autonomía, infieren resultados como predicciones, contenidos o decisiones que pueden influir en entornos físicos o virtuales (Unión Europea, 2024). En la práctica, la IA se apoya en diferentes enfoques, destacando:

- **Sistemas Expertos:** son programas que estructuran conocimientos especializados y reglas de inferencia para resolver problemas en un dominio específico, con el fin de apoyar la resolución de problemas en ese ámbito. Al depender de conocimiento previamente formalizado, estos sistemas pueden omitir criterios contextuales o valorativos que solo el ser humano logra percibir. En el campo jurídico, los Sistemas Expertos Jurídicos (SEJ) representan normas y conocimientos doctrinales en forma de reglas y permiten abordar problemas utilizando un tipo de razonamiento semejante al que realizaría un jurista experto en ámbitos específicos del Derecho (Torres, et al., 2022).

- **Aprendizaje Automático (*Machine Learning*, ML):** Es la capacidad de la IA de aprender de datos y acciones anteriores mediante algoritmos que buscan patrones para tomar decisiones (Barco, 2023). En el campo jurídico, ejemplos recientes en Colombia, tenemos a PretorIA, de la Corte Constitucional, diseñado para preclasificar y buscar tutelas mediante categorías y filtros; y PRISMA, de la Fiscalía,

que perfila riesgo de reincidencia a partir de datos de la Fiscalía, Policía e INPEC para sustentar solicitudes de medida de aseguramiento. Ambos ilustran el potencial y los dilemas de estas herramientas en Colombia (Torres, et al., 2022).

- **Aprendizaje Profundo (*Deep Learning*):** Emula redes neuronales complejas, extrayendo patrones de ingentes masas de datos masivos. Sus resultados no se vinculan de modo lineal, sino complejo, lo que puede llevar a la decisión algorítmica a ser opaca, conocida como "caja negra" o “black box” (Cancio, 2023). Un ejemplo sobre este sistema basado en IA lo ubicamos en Reino Unido, el caso ya mencionado de *R (Bridges) v Chief Constable of South Wales Police* (2020) sobre reconocimiento facial en vivo que evidenció una captación masiva no consentida, criterios opacos y riesgos de sesgos y falsos positivos, mostrando cómo el *deep learning* potencia la vigilancia a la vez que tensiona garantías de derechos fundamentales.

De otro lado, en el estudio de la Inteligencia Artificial suele distinguirse entre la llamada IA estrecha o débil y la IA general o fuerte, diferencia la cual resulta fundamental para comprender el alcance real de estas tecnologías**:**

1. **IA Estrecha o Débil:** Un sistema orientado a resolver problemas concretos y delimitados, que aprende a través de patrones repetitivos mediante algoritmos programados por humanos, y se restringe a un área específica de funcionamiento. Ejemplos prácticos son los asistentes virtuales o sistemas de reconocimiento facial (Hoz, K. y Coelho, F., 2021 citado en Vázquez, 2025).

2. **IA General o Fuerte:** El ideal de un sistema complejo capaz de emular la inteligencia humana en su amplitud, con comprensión, razonamiento y capacidad de abordar decisiones de forma proactiva, deductiva y autoconsciente (Hoz, K. y Coelho, F., 2021 citado en Vázquez, 2025).

La IA actual se caracteriza por tener una inteligencia de tipo refleja y no reflexiva, es decir, puede resolver tareas que calificamos como inteligentes, pero lo hace sin "saberlo" o reflexionar sobre ellas. En ese sentido, la IA General o Fuerte se mantiene aún en el plano de la proyección teórica y no ha logrado materializarse así, los sistemas de IA disponibles hoy corresponden a formas de IA estrecha, centradas en tareas específicas, aunque en muchos casos alcancen un grado de sofisticación muy elevado (Innerarity. 2025).

4.1.1.2 Inteligencia artificial en la toma de decisiones y el control estatal.

La distinción anterior permite pasar del plano conceptual al análisis de experiencias concretas de incorporación de IA en dispositivos estatales de decisión y control. En Colombia, sistemas como PretorIA en la Corte Constitucional (Solar, 2020) y PRISMA en la Fiscalía (López et al., 2023) muestran que, aun tratándose de IA esencialmente estrecha —orientada al triaje de tutelas o a la estimación de riesgo de reincidencia—, su uso incide directamente en derechos fundamentales, en la selección de casos y en la adopción de medidas restrictivas de la libertad, lo que exige altos estándares de transparencia y control humano significativo.

De forma paralela, el contraste con experiencias internacionales como COMPAS en EE. UU. (Ibáñez, 2023) o los sistemas de puntuación social proscritos en Europa —entre ellos SyRI en Países Bajos— (Presno, 2025; Pérez, 2025), permite evidenciar que el

problema no es solo técnico, sino constitucional: allí donde la decisión algorítmica afecta personas, surge la necesidad de evaluar sesgos, discriminación indirecta y legitimidad democrática de estos sistemas (Balaguer Callejón, 2025).

Con ello, queda más que claro que la irrupción de los sistemas de Inteligencia Artificial (IA) en la gestión pública y en los dispositivos de poder, vigilancia y control estatal ha generado una reconfiguración de las dinámicas sociales y jurídicas contemporáneas. En coherencia se realizará un recorrido analítico de carácter deductivo que parte de los grandes marcos de la teoría crítica del poder y del control social para, progresivamente, aterrizar en los conceptos operativos de la IA estatal, sus formas de intervención, los estándares exigibles en un Estado Social de Derecho y los principales marcos de gobernanza algorítmica pertinentes para el problema de investigación.

4.1.2. Marcos de análisis del poder punitivo en la era digital: la IA como dispositivo de poder y control estatal

El análisis de la IA en dispositivos estatales de poder y control requiere trascender el enfoque técnico-formal y recurrir a teorías que expliquen las dinámicas de poder subyacentes en las prácticas de vigilancia y gestión social (Pérez, 2023). Dicho en otras palabras, el análisis del uso estatal de la IA no puede entenderse solo como una herramienta neutra de optimización, sino que requiere comprenderla como un nuevo dispositivo de poder que se articula con las lógicas históricas de disciplina, vigilancia y control.

En esta línea foucaultiana, ello implica reconocer que la IA opera en el nivel material de las técnicas, los aparatos y las instituciones, reproduciendo principios de vigilancia, normalización y examen que constituyen el núcleo de las prácticas punitivas modernas. Así,

su estudio exige atender a la tecnología "en acción" como expresión del vínculo poder-conocimiento que organiza las formas contemporáneas de control (Garland, 1999).

4.1.2.1 Poder, vigilancia y biopolítica: del panóptico a la gubernamentalidad algorítmica

Las tradiciones sociológicas y filosóficas del castigo y el control social ofrecen herramientas críticas esenciales para interpretar el fenómeno de la IA estatal. David Garland (1999), acuño un uso distinto al termino "penalidad", tradicionalmente reducido al "*entramado de leyes, procedimientos, discursos, representaciones e instituciones que integran el ámbito penal*" (1999, p. 33), para ampliarlo y referirse con él al conjunto de prácticas, saberes y relaciones de poder que configuran el castigo en un sentido más amplio: no solo como un acto jurídico, sino como un fenómeno cultural, político y social (Garland, 1999); Y, siguiendo a Foucault (2002), ese poder no opera solo desde arriba, sino que se materializa en microformas de vigilancia y normalización —como el panóptico, los registros disciplinarios, las evaluaciones constantes o la comparación de sujetos— que producen comportamientos dóciles.

Ese mismo principio se actualiza hoy en el mundo digital: así como el panóptico hacía visible al interno para corregirlo (Foucault, 2002), en la actualidad, se ha configurado "*un nuevo panóptico, ahora digital y sin vigilantes*" (Barrios, 2025, p. 84), que gracias a los sistemas algorítmicos hacen visibles a los ciudadanos mediante datos, los clasifican, los puntúan y los ordenan según su "riesgo" (Hernández, 2025), dando lugar a lo que puede llamarse una "*gubernamentalidad algorítmica*" (Tirado Serrano et al, 2025, p. 2; Innerarity, 2025, p. 380), es decir, un modo de gobernar a través de información, predicción y segmentación de poblaciones en donde el control y decisión una es escasa o insuficiente.

De este modo, la IA estatal no aparece como un instrumento neutro, sino como la versión contemporánea de esos dispositivos de vigilancia, control y poder que Foucault describió. Con lo anterior, la sociología del castigo se concentra en las tecnologías reales del poder y su funcionamiento interno, analizando los principios de vigilancia y disciplina inscritos en las diversas instituciones que componen la sociedad (Garland, 1999). El concepto de "*continuum carcelario*" propuesto por Foucault (como se citó en Garland, 1999, p. 282-283) describe cómo las técnicas disciplinarias difunden los principios de identificación de transgresiones, anomalías y desviación de las normas a lo largo de todo el cuerpo social, abarcando desde la escuela y la familia hasta la prisión. Este marco de vigilancia y corrección se aplica a todo, desde la mínima irregularidad hasta el crimen atroz. De ahí la noción de microfísicas del poder: formas sutiles, cotidianas y técnicas de control que operan en escuelas, hospitales, cuarteles, prisiones o fábricas, y que producen sujetos obedientes mediante vigilancia, registro y normalización (Foucault, 1979).

En ese contexto, siguiendo la lectura foucaultiana que presenta Garland (1999), el castigo muestra que el poder no actúa solo reprimiendo, sino produciendo sujetos. En la "microfísica del poder", este se ejerce en los puntos de contacto más concretos con el transgresor. Técnicas como el examen y el registro minucioso de casos permiten observar, comparar y clasificar a las personas, generando sobre ellas un saber detallado y sistemático (Garland, 1999, p. 176). Así, poder y conocimiento se entrelazan: al describir al individuo, al mismo tiempo lo moldean y lo convierten en objeto de control.

A partir de lo anterior, la IA en el ámbito jurídico puede analizarse a través de la lente foucaultiana, en donde el castigo representa una "*táctica política situada en el campo general de las relaciones de poder*" (Garland, 1999, p. 166) que da origen a nuevas racionalidades y

técnicas para dirigir la conducta, del mismo modo en que Foucault lo mostró en su análisis de la prisión. En esta línea, los estudios contemporáneos sobre penalidad y control social —aunque más explícitamente sociológicos— prolongan la actitud crítica de Foucault al invitarnos a desnaturalizar el orden vigente e identificar los perjuicios, exclusiones y asimetrías de poder que se ocultan tras formas aparentemente neutrales de gestión de la conducta, hoy también mediadas por sistemas de IA.

La tecnología disciplinaria se concentra en el control meticuloso del cuerpo y la conducta (Foucault, 1979). En la era digital, esta lógica se traduce en la racionalización administrativa y la gestión mediante tecnologías que buscan incrementar la eficiencia y el control (Innerarity. 2025). Con ello, el trabajo de Foucault —y su relectura por autores como Garland (2001)— inspira las preguntas sobre la función que cumplen las prácticas penales en el gobierno de la sociedad tardomoderna; de manera análoga, autores recientes como Innerarity (2025) desplazan esas mismas inquietudes hacia nuestro tiempo convulso, en el que la racionalización y el control se ejercen crecientemente a través de sistemas algorítmicos y tecnologías de IA.

A partir de estas bases, el análisis se extiende a la biopolítica y la gubernamentalidad. La biopolítica "*se refiere a las estrategias de gobierno involucradas con la vida, la salud,* la *eficiencia y la seguridad de toda la población*" (Foucault, 1977 citado en Garland, 1999, p. 163). Por su parte, la gubernamentalidad en su sentido clásico refiere a la manera en que se ejerce el poder para dirigir las conductas de los individuos y las poblaciones. En la era digital, esta noción se reformula en términos de "*gubernamentalidad algorítmica*" (Castro Serrano et al, 2025), que surge como un marco para examinar cómo el poder se organiza y circula en sociedades crecientemente mediadas por algoritmos

Daniel Innerarity retoma estas preocupaciones al elaborar, en *Crítica de la razón algorítmica*, una filosofía política de la inteligencia artificial. Su teoría crítica entiende la "gubernamentalidad algorítmica" como una forma de poder que naturaliza la automatización y oculta la no neutralidad de las decisiones inscritas en los datos y los modelos. Frente a la ideología de una racionalidad técnica sin alternativas, Innerarity insiste en preguntar qué lugar conserva la decisión política y el control democrático sobre tales sistemas.

Esta perspectiva permite traducir la biopolítica y la gubernamentalidad foucaultianas al presente digital y abre el paso a enfoques críticos que subrayan las implicaciones de estas tecnologías para la vigilancia y la producción de nuevas formas de docilidad social. En ese orden de ideas, autores como Shoshana Zuboff (2019) han ampliado el concepto de vigilancia, advirtiendo sobre la vigilancia biométrica masiva que se presenta de forma seductora e invisible, llevando a la ciudadanía a convertirse en una "masa dócil pastoreada por algoritmos que giran en torno a sus datos" (Balaguer & Escajedo, 2025, contraportada).

En este tipo de vigilancia seductora e invisible, el interés no se centra en el Estado como aparato represivo, "*sino en cómo este se gubernamentaliza mediante procesos e instrumentos técnico-discursivos que articulan marcos de inteligibilidad y modos de acción acordes con el capitalismo contemporáneo*" (Castro Serrano et al., 2025, p. 231).

Es decir, el foco se desplaza del Estado Leviatán despótico, hacia las redes de saberes, normas y dispositivos que definen qué es pensable y qué acciones son legítimas. Esos instrumentos técnico-discursivos producen formas de ver el mundo y de orientar conductas funcionales al capitalismo contemporáneo —productividad, eficiencia, competencia— sin requerir siempre de la coerción directa. En este marco, la algoritmización de las decisiones

políticas y sociales obliga a repensar categorías fundamentales como sujeto, acción, responsabilidad y conocimiento (Innerarity. 2025).

De todo ello, se logra entender que la sociología del castigo ha experimentado importantes desarrollos cualitativos, pasando de una fase descriptiva (identificación de tendencias penales) a una de comprensión interpretativa, explicación causal y desarrollo teórico más analítico (Garland, 2019). Los estudios recientes operan en un nivel menos abstracto que los marcos teóricos generales de Marx, Durkheim, Weber o Foucault, enfocándose en efectos y casos específicos, que aterrizados a nuestros días presentan una enorme preocupación sobre el uso e influencia de la IA en las nuevas lógicas de poder, disciplina y control estatales.

4.1.2.2 Gubernamentalidad algorítmica y capitalismo de la vigilancia como lógicas contemporáneas de poder

El auge de la IA plantea la necesidad de nuevas herramientas conceptuales para analizar procesos de poder y control, como la ya mencionada Gubernamentalidad Algorítmica (Castro Serrano et al, 2025), Estos sistemas de IA se han configurado con el potencial de invadir la interioridad humana y reconfigurar las subjetividades en sus dinámicas sociales, éticas e identitarias (Barrios, 2025, citado en Balaguer & Escajedo, 2025).

Shoshana Zuboff describe *el Capitalismo de la Vigilancia* (2019) como una lógica de acumulación sin precedentes, definida por nuevos imperativos económicos. Este sistema funciona capturando y registrando de manera continua los rastros digitales que dejan las personas al usar plataformas y dispositivos. A partir de ahí no solo se recoge la información

necesaria para prestar el servicio, sino también un **"***excedente conductual***"** (Zuboff, 2019, p. 237): datos adicionales sobre hábitos, emociones y preferencias que se extraen y procesan sin que el usuario lo perciba.

Ese excedente se transforma, mediante técnicas de análisis masivo y aprendizaje automático, en modelos capaces de anticipar comportamientos futuros, que se comercializan como productos predictivos altamente rentables. Esto genera una "división patológica del aprendizaje" (Zuboff, 2019, pp. 231, 543), en la que la capacidad de aprender de la experiencia y convertirla en conocimiento útil se concentra en una élite decisora y en las infraestructuras algorítmicas que opera.

Esta lógica se manifiesta en era de la *algocracia* (Aneesh, 2009; Danaher 2016, citado en Innerarity. 2025, p. 314), un sistema que emplea algoritmos para recoger, cotejar y organizar los datos a partir de los cuales se toman las decisiones, de manera tal que IA adopta el papel de burócratas y creadores de mitos (Harari, 2024) que buscan modificar y optimizar el comportamiento humano, a menudo sin que el individuo sea consciente de ello para definir la cuestión fundamental de quien o que ejerce "*el conocimiento, la autoridad y el poder en nuestro tiempo: quién sabe, quién decide, quién decide quién decide" (*Zuboff, 2019, p. *224)*.

4.1.2.3 Formas de intervención estatal mediadas por inteligencia artificial: perfilamiento, priorización y predicción

Conforme a lo anterior, el fenómeno de la IA como dispositivo estatal se estudia a partir de su definición operativa, sus tipologías y los modos específicos en que interviene en las funciones de poder, vigilancia y control. En la esfera pública, estos sistemas se incorporan principalmente para atender problemas de descongestión y demoras judiciales, automatizar

tareas repetitivas y asistir en la toma de decisiones administrativas o jurisdiccionales (Fonseca, 2022).

Sin embargo, no solo optimizan procesos, sino que introducen nuevas formas de visibilizar, clasificar y jerarquizar a las personas. Por ello resulta imperativo analizar los tres modos en que os algoritmos intervienen como dispositivos de poder, vigilancia y control:

1. **Perfilamiento (Profiling):** La IA utiliza datos para crear perfiles de individuos o grupos, lo que se relaciona estrechamente con la prohibición de discriminación y la vigilancia (Jiménez, 2025). En el ámbito penal, el perfilamiento para tomar decisiones que recaen sobre personas individualizadas, lo que puede afectar su presunción de inocencia (Solar, 2022, como se cita en Hernández 2025, p. 47 – 48). Otros ejemplos son los sistemas de IA utilizados para la puntuación ciudadana (que evalúan o clasifican personas en base a su comportamiento social o características de personalidad), los cuales han sido considerados como un riesgo inaceptable en la Unión Europea (2024), ya que pueden tener resultados discriminatorios y menoscabar la dignidad de los evaluados (Escajedo, 2025).

2. **Predicción:** La IA se utiliza en funciones públicas para anticipar comportamientos o resultados futuros, mediante modelos que estiman probabilidades a partir del análisis de datos históricos y variables seleccionadas (Londoño, J. 2024). Un ejemplo paradigmático es la predicción del riesgo de reincidencia criminal (Sánchez, 2021) como COMPAS en Estados Unidos y PRISMA en Colombia. Sin embargo, cuando estos perfiles predictivos se emplean para orientar o sustituir la decisión judicial, pueden afectar el derecho del acusado a un juicio imparcial y al principio de igualdad (Solar, 2022, como se cita en Hernández 2025, p. 49).

3. **Priorización:** Los sistemas de IA en la administración pública pueden utilizarse para seleccionar beneficiarios de servicios, asignar becas o determinar la elegibilidad para beneficios penitenciarios (Solar, 2020). La priorización algorítmica tiene efectos distributivos relevantes, en la medida en que decide quién accede a recursos, beneficios o tratos favorables y quién queda excluido o soporta mayores cargas, lo cual debe ser contrastado y analizado a la luz de la igualdad material y la dignidad. Un antecedente ilustrativo es SyRI (Países Bajos), que cruzaba masivamente datos para perfilar fraude en prestaciones y que fue anulado por el Tribunal de La Haya por opacidad y afectación desproporcionada de la vida privada y discriminación, evidenciando dichos efectos distributivos negativos.

Con todo ello, la incorporación de IA en decisiones estatales también puede derivar en formalismo y reduccionismo. Y es que, al traducir los conflictos humanos a problemas de predicción, clasificación o puntuación, la IA tiende a privilegiar la llamada *racionalidad objetiva*: aquella que se somete al criterio lógico-formal del modelo y deja en segundo plano la finalidad humana y social de la decisión (Gómez et al., 2025).

En el ámbito del control y del castigo de nuestra algocracia, se corre el riesgo de convertir a la justicia en una operación matematizada, donde lo que pesa no es la singularidad del caso ni el contexto del sujeto, sino el ajuste del dato al modelo. Así, la herramienta que debía apoyar al juez o al fiscal puede terminar instrumentalizando a la persona y desplazando los criterios de utilidad social, equidad y dignidad que el Estado Social de Derecho exige mantener en el centro.

4.1.3. Inteligencia artificial y Estado Social de Derecho: tensiones sobre los derechos fundamentales

La implementación de la IA en dispositivos estatales exige una lectura crítica y garantista de los sistemas a la luz de los principios constitucionales, especialmente aquellos que definen el Estado Social de Derecho (ESDD). Así las cosas, el punto de partida es considerar que el problema de la IA en el Estado no radica en la tecnología en sí misma, sino en sus efectos colaterales sobre los derechos fundamentales según la forma en que se diseña, se entrena y, sobre todo, como se emplea (Salvador, 2025). En ese contexto, el auge de la IA exige reflexionar sobre la transformación o replanteamiento de las categorías jurídicas, particularmente los derechos fundamentales (Sarrión, 2023).

4.1.3.1 El Estado Social de Derecho colombiano ante la IA: reconfiguración de las garantías fundamentales

En Colombia, el Estado Social de Derecho tiene como fines esenciales garantizar la efectividad de los principios, derechos y deberes, promover la prosperidad general y asegurar un orden justo (Younes, 2021). La expansión acelerada de la IA en la sociedad y, en particular, en la administración pública —incluido el sistema judicial— no solo introduce nuevas herramientas de gestión, sino que reconfigura las condiciones en las que se ejercen y protegen los derechos fundamentales. Más que un simple déficit regulatorio técnico, este proceso plantea la necesidad de repensar, en a la luz de la constitución, cómo orientar la implementación de la IA de modo compatible con los fines y garantías propios del Estado Social de Derecho (Fonseca, 2024).

En ese contexto, la IA no es neutral respecto a los derechos fundamentales; los impacta directa y complejamente. Por ello, la simple extensión mecánica de los derechos ya existentes resulta insuficiente. Surgen así debates sobre la necesidad de reconocer nuevos derechos o, al menos, nuevas facetas de los ya consagrados (Roig, 2024), como el derecho al control humano sobre los sistemas de IA, fin de asegurar una tutela efectiva de los bienes jurídicos que el Estado se compromete a proteger (Gómez et al., 2025; Hernández, et al 2025).

En este marco, la investigación se centra en examinar cómo la incorporación de sistemas de inteligencia artificial en los dispositivos estatales de poder, vigilancia y control reconfigura tres garantías fundamentales del Estado Social de Derecho: la dignidad humana, la igualdad material y la prohibición de discriminación.

A. Igualdad material y no discriminación frente a la IA estatal

En el Estado Social de Derecho, el derecho a la igualdad ocupa un lugar cardinal, en cuanto exige que todas las personas sean tratadas de manera justa y no discriminatoria (Younes, 2021).

La discriminación algorítmica es uno de los mayores riesgos de la IA, la cual se manifiesta cuando la toma de decisiones automatizada perpetúa sesgos y desigualdades preexistentes o formula unos nuevos (Coddou et al., 2025). Esta preocupación no es solo académica, incluso la Política Nacional de Inteligencia Artificial, formulada mediante el Documento CONPES 4144 (2024), reconoce la necesidad de identificar, prevenir y mitigar los riesgos y efectos no deseados de los sistemas de IA para evitar asimetrías, inequidades y posibles vulneraciones de derechos en el país.

En otras palabras, la propia política pública admite que, si no se gobierna adecuadamente, la IA puede convertirse en un mecanismo que profundiza, amplia y perpetua los prejuicios sociales existentes, reproduciendo dinámicas de poder arraigadas en las relaciones sociales en nuevos escenarios digitales (Serrano et al, 2025). Ejemplos de esta perpetuación incluyen la utilización de IA en la persecución del crimen, la educación y el acceso al mercado laboral[10].

En este contexto, la igualdad material, entendida como un trato orientado a garantizar condiciones reales y efectivas de protección, más allá de la mera igualdad formal, requiere según Sánchez (2021) que los sistemas de IA:

1. Respeten el acceso a las mismas oportunidades, considerando las condiciones socioeconómicas y particulares de cada persona (tratar igual a los iguales y desigual a los desiguales).

2. Eviten que las decisiones algorítmicas afecten desproporcionadamente los derechos de las personas.

Por lo tanto, surge la necesidad de ir más allá del *formalismo* o la *trampa* al mismo, lo cual implica que la equidad no se reduzca a métricas abstractas sin considerar el contexto

[10] Un caso relevante en el acceso al empleo es el sistema experimental de reclutamiento de Amazon. Entre 2014 y 2017, la empresa entrenó un modelo de aprendizaje automático para puntuar hojas de vida de una a cinco estrellas a partir de historiales de contratación previos; como esos datos estaban dominados por hombres, el sistema aprendió a penalizar currículos de mujeres —por ejemplo, aquellos que mencionaban la pertenencia a grupos femeninos como “capitana del club de ajedrez femenino"— y a favorecer sistemáticamente candidatos masculinos, razón por la cual fue finalmente descartado (Dastin, 2018). Estudios posteriores han mostrado que este diseño no solo reproducía, sino que consolidaba la desigualdad estructural: el algoritmo operó como un filtro que reducía significativamente las tasas de contratación de mujeres en tecnología de la información y evidenciaba la necesidad de estrategias específicas para mitigar el sesgo de género en la selección automatizada (Chang, 2023).

social y político de quien es estudiado o analizado, con lo cual es crucial que el análisis técnico de sesgos se guíe por el contexto del problema y las consecuencias sociales, seleccionando métricas de equidad —criterios cuantitativos que se usan para medir si un sistema de IA trata de forma justa a distintos grupos de personas independientemente a su género, raza, nivel socioeconómico, territorio, etc.—, que permitan aproximarse a los estándares del derecho antidiscriminación, por ejemplo, en el ámbito penal, donde la libertad está en riesgo, se debe priorizar la minimización de falsos positivos (Coddou et al., 2025).

Ahora bien, la sola proclamación de la igualdad no basta cuando las decisiones automatizadas se entrenan con datos que reflejan un pasado desigual. "*Los sistemas de IA se diseñan para alcanzar objetivos y métricas de rendimiento específicos (precisión, recall, paridad demográfica)"* (Coddou et al., 2025, p. 90), por lo tanto, la discriminación algorítmica y la reproducción de desigualdades surgen porque los datos nunca son completamente neutrales: incorporan trayectorias históricas de exclusión que el modelo termina aprendiendo (Innerarity. 2025).

En este escenario, los sistemas de IA reproducen las desigualdades en tres pasos: primero, las desventajas sociales se plasman en los datos; luego, esa realidad se refuerza en la norma o en el modelo como si fuera una "situación objetiva"; y, finalmente, se consolida mediante decisiones posteriores que toman esos resultados como válidos (Innerarity. 2025, p. 356). Una primera respuesta a este problema consiste en exigir que los algoritmos se entrenen con grandes volúmenes de datos de calidad, suficientemente representativos de los distintos grupos de la población, de lo contrario, el sesgo presente en la muestra de entrenamiento se incorpora al sistema como un criterio más de selección o clasificación (Fernández, 2019 citado en Hernández, 2025).

Sin embargo, el problema de los sesgos no se agota en la fase de diseño o entrenamiento. Pueden surgir en cualquiera de las etapas del ciclo de vida del sistema: en su uso y aplicación, en los procesos de validación o incluso en la forma en que se presentan los resultados. Entre los tipos de sesgos identificados se encuentran los errores de medida, los sesgos culturales, los estereotipos de lenguaje y los sesgos de exclusión, es decir, aquellos que omiten sistemáticamente a una parte de la población en la muestra (Hernández, 2025).

cuando los algoritmos toman el pasado como patrón para decidir sobre el futuro, se configura un sistema que perpetúa trayectorias de desigualdad de las que resulta difícil escapar, aumentando la dependencia a dicha trayectoria y reduciendo la posibilidad de elegir libremente (Innerarity. 2025). Esta lógica se manifiesta, por ejemplo, en herramientas como COMPAS, PRISMA y el Fiscal Watson, donde los sesgos presentes en los datos históricos tienden a reproducirse en la evaluación del riesgo penal y penitenciario.

En términos concretos, el caso de COMPAS mostró que, pese a tasas similares de reincidencia, el sistema clasificaba con mayor frecuencia a personas negras como de “alto riesgo”, reproduciendo patrones policiales discriminatorios. En Colombia, PRISMA, al entrenarse con capturas y antecedentes registrados en el SPOA, tendía a sobrerrepresentar barrios y grupos sometidos a una nueva hipervigilancia, reforzando trayectorias de criminalización y exclusión. De forma análoga, el Fiscal Watson, al priorizar casos y sugerir líneas de investigación a partir de grandes bases documentales, corre el riesgo de amplificar estereotipos y omitir contextos relevantes, de modo que las decisiones futuras se apoyen en mapas delictivos ya sesgados y consoliden la desigualdad.

B. Dignidad humana y autonomía como fundamentos constitucionales del control humano frente a la IA

El análisis del poder y el control debe anclarse en la defensa de la dignidad humana como el valor y principio supremo del ordenamiento jurídico colombiano Younes, 2021). La IA, en su uso estatal, amenaza la dignidad cuando el ser humano es instrumentalizado como un medio para alcanzar otros fines, en lugar de ser concebido como un fin en sí mismo.

La dignidad es inherente a todas las personas y constituye la base de los derechos humanos, fundamentando la consideración de las personas como sujetos libres e iguales. En el contexto de la IA, la dignidad funciona como un límite infranqueable para el uso de sistemas inteligentes. Esto implica que la persona no puede ser objetivada ni instrumentalizada, incluso en escenarios artificiales (Galera, 2025).

La autonomía humana, entendida como la capacidad de determinar el propio destino y desarrollar la personalidad en libertad y con respeto a la dignidad, se encuentra bajo amenaza cuando las decisiones humanas son influenciadas por sistemas automatizados opacos. El objetivo último de la IA debe ser aumentar el bienestar humano y mejorar las capacidades humanas, y nunca sustituir o poner en peligro la dignidad (Galán, 2023).

En ese contexto, la IA debe estar sujeta a la primacía de la dignidad humana (Hernández, 2025). Este mandato implica que no basta con garantizar un correcto funcionamiento técnico de los sistemas, sino que es preciso asegurar que las decisiones automatizadas permanezcan bajo la órbita de responsabilidad y revisión humana. El fundamento del control humano como derecho radica en su vínculo intrínseco con la dignidad humana: el control humano opera como la respuesta jurídica que permite extender la exigencia de respeto por la dignidad a los escenarios mediados por sistemas de IA.

Consecuentemente, el principio de dignidad se proyecta en la autonomía individual, en la medida en que exige que las personas conserven un margen efectivo de control sobre la aplicación y el impacto de los sistemas inteligentes en sus vidas, de modo que estos funcionen como instrumentos para potenciar su autonomía y no como mecanismos de intromisión o sustitución de su capacidad de decisión.

Así las cosas, los procesos algorítmicos pueden debilitar la autonomía cognitiva de los individuos y su derecho a formar opiniones y tomar decisiones independientes, elementos centrales del Estado de Derecho (Unión Europea, 2024). Esta preocupación por la autonomía cognitiva y la protección de la esfera mental conduce, en algunos contextos, a la formulación de nuevos derechos los llamados neuroderechos.

C. Neuroderechos: protección de la identidad, la integridad mental y el libre albedrío frente a la IA

El impacto de las neurotecnologías y la IA ha llevado a la postulación de los neuroderechos**, "***entendidos como aquellos que protegen a las personas de los eventuales peligros por los efectos de las neurotecnologías, que se equiparan a cualquier herramienta o técnica capaz de manipular, registrar, medir y obtener información del cerebro*" (Hernández et al et al., 2025, p. 54). En esta línea, los neuroderechos apuntan a salvaguardar la esfera mental frente a riesgos como la manipulación o explotación de la información cerebral, que podrían vulnerar la privacidad mental, la identidad personal, la integridad psicológica y el libre albedrío (Hernández et al., 2025). Un ejemplo paradigmático es Chile,

que ha reformado su Constitución para reconocer y proteger la actividad cerebral como un ámbito especialmente reforzado de derechos[11].

En esta perspectiva, los neuroderechos no sólo buscan proteger el cerebro como órgano, sino resguardar la propia experiencia subjetiva frente a tecnologías capaces de leer, inferir o modificar estados mentales. En particular, la IA tiene el potencial de configurar subjetividades y reconfigurar la identidad de los individuos (Barrios, 2025). Desde ahí, la dignidad humana opera como límite a la aplicación de estos sistemas, en tanto representa el valor central que engloba la autonomía, el derecho a existir y a elegir (Galera, 2025). La vulneración de la dignidad humana se convierte así en una preocupación principal, especialmente cuando la clasificación automatizada adquiere rasgos cosificantes o estigmatizantes.

Ahora bien, frente a estos dilemas del uso de la IA en decisiones estatales surge la exigencia de control humano significativo: toda herramienta algorítmica que pueda afectar derechos debe estar sometida a la supervisión, validación y eventual corrección de una autoridad humana responsable. Esto implica que el sistema no puede decidir de forma autónoma y cerrada, sino que su resultado debe ser comprensible, discutible y revisable por el operador jurídico, manteniendo así la centralidad de la persona y del juicio práctico en el Estado Social de Derecho.

[11] Esta reforma, surgida de la preocupación ante los riesgos que plantea el acceso y la posible manipulación de la información cerebral mediante el uso de neurotecnologías como el interfaz cerebro-computador, se materializó con la Ley Nº 21.383 del 25 de octubre de 2021. Dicha ley modificó el artículo 19 Nº1 de la Carta Magna para asegurar que el desarrollo científico y tecnológico esté al servicio de las personas y, de manera crucial, que la ley deba "*resguardar especialmente la actividad cerebral, así como la información proveniente de ella*", protegiendo así aspectos fundamentales del individuo como la privacidad mental, la identidad personal, la integridad psicológica y el libre albedrío (Hernández, 2025).

4.1.4. El Control Humano Significativo como derecho y mecanismo instrumental de garantía preventiva y correctiva

Al operar la IA con cierto grado de autonomía, tiende a desplazar la intervención directa de la inteligencia humana y a ocultar al usuario las etapas del proceso decisorio (Quiceno, 2025). Esta opacidad no es solo un rasgo técnico, sino un problema jurídico, porque dificulta reconstruir quién decide realmente y con base en qué criterios lo ha hecho. De ahí que cobre relevancia el control humano significativo como derecho: se requieren herramientas algorítmicas y salvaguardas institucionales que garanticen una deliberación humana efectiva sobre los resultados del sistema y permitan corregirlos cuando se muestren incompatibles con los derechos fundamentales (Flórez, 2024)

El Control Humano Significativo (CHS)[12] es un concepto clave que esta investigación debe definir y distinguir entre sus funciones preventiva y correctiva, para usarlo como criterio de evaluación de la gobernanza de la IA estatal (Cavalcante Siebert et al., 2023). Desde esta perspectiva, la supervisión humana no puede reducirse a una presencia meramente formal —un operador pasivo que se limita a ratificar lo que la máquina ya decidió—, sino que debe implicar capacidad real de cuestionar, modular o revertir el resultado algorítmico (Santoni de Sio y van den Hoven, 2018). En últimas, la responsabilidad por las decisiones asistidas o mediadas por IA recae en agentes humanos que diseñan, parametrizan, implementan y utilizan estos sistemas, de modo que el CHS se convierte en una condición para que la atribución de responsabilidad y la exigencia de rendición de

[12] En lo sucesivo, la sigla CHS se empleará para referirse al Control humano Significativo

cuentas sigan siendo posibles materialmente (Davidovic, 2023; Santoni de Sio & Van den Hoven, 2018).

Conceptualmente, aunque el control humano atraviesa transversalmente el análisis de la gubernamentalidad algorítmica y de las vías para evitar que los sistemas de IA lesionen derechos fundamentales, la expresión "Control Humano Significativo" (CHS) —*meaningful human control* (MHC) en la literatura anglófona— tiene un origen más acotado al debate del Derecho Internacional Humanitario sobre los sistemas de armas autónomos letales —LAWS por su sigla en inglés—, plataformas militares que puede identificar, seleccionar y atacar objetivos de forma autónoma, sin intervención humana durante el ciclo de ataque (Santoni de Sio y van den Hoven, 2018 ; Cavalcante Siebert et al., 2023).

En ese contexto bélico se formuló el CHS como exigencia de que exista siempre una persona con capacidad real de supervisar, intervenir y asumir responsabilidad por el uso de la fuerza, frente al riesgo de armas que operen sin suficiente control humano (Santoni de Sio y van den Hoven, 2018). Con el tiempo este vocabulario se ha extendido al campo más amplio de la gobernanza de la IA y de los sistemas automatizados de decisión en ámbitos civiles —incluidos la justicia penal, la seguridad social o las políticas públicas—, donde el CHS se presenta como una herramienta clave para asegurar seguridad, dignidad y responsabilidad institucional frente a decisiones algorítmicas de alto impacto (Davidovic, 2023).

Así las cosas, puede definirse el CHS como la capacidad efectiva de intervención de los seres humanos a lo largo de todo el ciclo de vida del sistema de IA —desde el diseño y entrenamiento hasta su despliegue, supervisión y eventual desactivación— con el objetivo de prevenir o mitigar impactos negativos sobre los derechos humanos. Esta capacidad supone

que los operadores comprendan el sistema en un grado suficiente, dispongan de información y explicaciones relevantes y cuenten con márgenes temporales y organizacionales para ejercer un juicio propio, en lugar de limitarse a convalidar la salida algorítmica, es decir, avalar o no un resultado establecido por la IA, lo cual se considera una supervisión meramente formal (Santoni de Sio & Van den Hoven, 2018; Siebert et al., 2023; Davidovic, 2023).

En el plano jurídico, una parte de la doctrina ha comenzado a formular el control humano —especialmente en contextos de decisión automatizada— como un nuevo derecho intrínsecamente relacionado a la vertiginosa expiación de tecnólogas algorítmicas basadas en IA. Se trata de un derecho de carácter instrumental toda vez que no tiene un contenido completamente autónomo, sino que se justifica por su función de garantía frente a otros derechos fundamentales, tales como la igualdad, la libertad, la privacidad y el debido proceso, cuya protección se ve comprometida cuando las decisiones públicas se delegan a sistemas opacos y por consecuencia incomprendidos (Sánchez, 2021; Pasquale, 2021).

Desde esta perspectiva, el CHS opera como condición de posibilidad para que esos derechos sigan siendo exigibles en entornos gobernados por algoritmos, especialmente cuando se trata de decisiones estatales de vigilancia, control y castigo. Sin embargo, el CHS no se agota en la mera posibilidad abstracta de intervención, sino que comprende diversos momentos de supervisión, participación, revisión y determinación humana (Santoni de Sio y van den Hoven, 2018), para así lograr materializar el tipo y grado de control que preserve la acción humana y sostenga la responsabilidad moral de estos en las decisiones (CICR, 2018).

En concreto, Sánchez (2021) agrupo las funciones esenciales del CHS en cuatro ejes: (i) la supervisión y veeduría humanas en el diseño y entrenamiento del sistema; (ii) la

disponibilidad y presencia humana durante las fases de desarrollo, despliegue y operación del sistema de IA; (iii) la revisión humana ex post para corregir, reparar o compensar eventuales afectaciones; y (iv) la facultad de apartarse de la decisión automatizada, ya sea desactivando el sistema, sustituyendo el resultado o absteniéndose de utilizarlo en determinados contextos.

La presencia de un CHS en los ejes mencionados permitirá diseñar arreglos institucionales que mantengan a la IA como medio al servicio de las personas y de las finalidades constitucionales, y no como un fin en sí mismo, preservando la autonomía humana y la capacidad efectiva de contestar y reorientar las decisiones algorítmicas.

Finalmente, el CHS exige un punto de equilibrio: debe ser lo suficientemente robusto para proteger derechos y asegurar responsabilidad, pero sin convertirse en un corsé tan rígido que haga inviable o puramente simbólico el uso de la IA (Sánchez (2021). Tal como subrayan las discusiones filosóficas recientes, definiciones excesivamente vagas del CHS tienden a vaciar el concepto y a legitimar formas de supervisión meramente aparentes, mientras que definiciones excesivamente fuertes corren el riesgo de bloquear innovaciones tecnológicamente útiles o socialmente valiosas (Davidovic, 2023; Cavalcante Siebert et al., 2023).

4.1.4.1 Criterios para la implementación del Control Humano Significativo

Ahora bien, para que ese Control Humano Significativo no se quede en una declaración abstracta, sino que opere realmente sobre los sistemas algorítmicos, es necesario traducirlo en condiciones técnicas y jurídicas verificables. Dicho de otro modo: si el humano ha de poder intervenir, revisar, corregir o incluso apartarse de una decisión automatizada,

entonces el sistema debe ser transparente, explicable, trazable y sometido a rendición de cuentas. Son estos componentes operativos los que permiten al operador conocer cómo se obtuvo un resultado, contradecirlo cuando sea necesario y atribuir responsabilidad al órgano público que lo usa, cerrando así el círculo preventivo y correctivo del CHS.

Estos componentes se describen a continuación:

- **Transparencia y Explicabilidad:** La IA se enfrenta al problema de la opacidad algorítmica. Cuando no existe claridad sobre la lógica del modelo, su código fuente o los datos de referencia, se impide conocer las razones de los resultados que ofrece el sistema y se comprometen diversos derechos fundamentales, al no ser posible entender el *porqué* de los resultados proporcionados (Pernas, 2023). Esta falta de transparencia refuerza la necesidad de exigir explicaciones comprensibles y de articular mecanismos de control humano significativo sobre las decisiones asistidas por IA. Por su parte la transparencia es un principio constitucional que exige que el gobierno y las empresas sean abiertos y accesibles en su rendición de cuentas, de modo que los sistemas algorítmicos deben diseñarse y gobernarse conforme a este principio para evitar la opacidad en su uso (Pérez Conchillo, 2025). El control humano, ejercido mediante la revisión cualificada de las decisiones automatizadas, busca precisamente acceder a los factores, la lógica y las técnicas que generaron el resultado (Flórez, 2025), lo que permite reducir de manera significativa el riesgo de menoscabar derechos fundamentales tales como la intimidad, el debido proceso, la igualdad, la no discriminación y la protección de datos personales.

Un ejemplo ilustrativo de los riesgos derivados de la ausencia de control humano significativo es la reciente sentencia de tutela de la Sala de Casación Civil de la Corte Suprema de Justicia (2025), que dejó sin efectos una providencia sustentada en citas jurisprudenciales inexistentes generadas con apoyo de IA. La Corte recordó que la motivación judicial no puede descansar en contenidos opacos o no verificables y, retomando la T-323 de 2024 (Corte Constitucional), advirtió que las "alucinaciones" de estos sistemas vulneran el debido proceso cuando el juez no verifica rigurosamente la veracidad y fiabilidad de la información incorporada. De este modo, vincula la transparencia y la explicabilidad algorítmica con el deber de motivar decisiones en términos accesibles, verificables y sometidos a control humano.

- **Trazabilidad y Rendición de Cuentas:** La gobernanza algorítmica implica la responsabilidad de las actuaciones que la IA realiza. Los sistemas de IA que tienen un impacto significativo deben ser susceptibles de recurso y ser auditables. Se requieren controles o auditorías efectivas en cuanto al contenido y funcionamiento operacional de los algoritmos y los sesgos que contienen. Las auditorías pueden ser individuales, colectivas, nacionales, internacionales, públicas, privadas y/o mixtas (Gómez et al., 2025).

Ahora bien, para que esos criterios de implementación del control humano significativo no se queden en una exigencia meramente formal, es indispensable que el propio sistema pueda ser comprendido, auditado y, en caso necesario, controvertido por el operador humano. Dicho de otro modo: solo puede haber un control humano real sobre la decisión automatizada si la decisión es mínimamente entendible.

Aquí aparece una dificultad típica de los sistemas contemporáneos de IA, sobre todo aquellos basados en modelos complejos o de aprendizaje profundo: la opacidad total o “caja negra”. Esta opacidad no solo es un reto técnico, sino que tiene consecuencias jurídicas directas, porque afecta el derecho de defensa, el debido proceso y la posibilidad de asignar responsabilidades. Por eso, a continuación, es necesario examinar el problema de la caja negra y la explicabilidad como condición previa para que el control humano significativo pueda operar efectivamente.

4.1.4.2 El problema de la caja negra y la explicabilidad de los sistemas de IA

El *deep learning* y, en general, los modelos de alta complejidad producen resultados en los que la relación causal entre los datos de entrada y la decisión final es difícil de reconstruir. Esta dificultad desdibuja la transparencia del proceso y hace que la toma de decisiones automatizada se perciba como una “caja negra”, es decir, un sistema cuyo funcionamiento interno resulta opaco incluso para sus propios diseñadores, en el que solo son observables las entradas y las salidas, pero no el conjunto de operaciones intermedias que conducen al resultado (Cancio, 2023).

La opacidad algorítmica no es solo un problema técnico: plantea un desafío ético, jurídico y de confianza pública, porque impide saber si el sistema actuó con criterios compatibles con los derechos fundamentales (Flórez, 2025). Aunque ciertos marcos normativos como el europeo han introducido un "derecho de explicación" en relación con el proceso de toma de decisiones automatizado con el objeto de tener mayores garantías, este no implica que todo el sistema de IA sea completamente explicable, ni que la explicación sea siempre comprensible para el operario profesional o el sujeto afectado (Innerarity, 2925).

La brecha entre el éxito predictivo del modelo y su poca capacidad explicativa puede terminar relegando la justificación, que es precisamente lo que el derecho exige, a un segundo plano, primando en consecuencia los resultados (Gómez et al., 2025). En el contexto colombiano, la STC17832-2025 de la Sala de Casación Civil (2025) muestra cómo una decisión puede convertirse en "caja negra" aun sin modelos complejos: el tribunal apoyó la terminación del proceso en citas jurisprudenciales inexistentes generadas con apoyo de IA, que la Corte asocia a "*alucinaciones*" no advertidas.

Al no poder la autoridad judicial reconstruir las razones reales de su decisión, se vulnera el debido proceso y se erosiona la confianza en la administración de justicia. Por ello la Corte exige el desarrollo de una alta diligencia en la verificación de la información empleada para motivar las providencias y advierte que la admisión acrítica de textos producidos por IA conduce a una motivación aparente, incompatible con la transparencia y la explicabilidad que el Estado Social de Derecho exige.

El caso resuelto en la STC17832-2025 muestra que el problema de la "caja negra" no se limita a los modelos complejos, sino a cualquier decisión en la que el juez pierde control sobre las razones efectivas del fallo. Para evitar que situaciones como esta se repitan, se ha propuesto "abrir la caja" mediante auditorías al código y a los datos, la revisión del control humano en cada etapa del ciclo de vida de la IA y, especialmente, la socialización del lenguaje algorítmico para operadores jurídicos y ciudadanía (Hernández, 2025).

Esta exigencia de "abrir la caja" **es** igualmente decisiva en los dispositivos de evaluación de riesgo penal y penitenciario, como COMPAS en Estados Unidos o PRISMA en Colombia. Cuando estos sistemas se incorporan a la decisión sin que sea posible conocer y discutir sus criterios, se ponen en cuestión garantías básicas del proceso: la persona

afectada carece de información para controvertir el insumo algorítmico, la defensa no puede cuestionar sus supuestos y el juez ve limitada su capacidad para ponderar su fiabilidad y peso probatorio. De este modo, se erosionan la imparcialidad y el carácter público y razonado de la decisión, pilares del debido proceso en el Estado Social de Derecho

En estos contextos, la exigencia de control humano significativo encuentra su justificación: la presencia de sesgos y la posibilidad de que la decisión se vuelva una "caja negra" justifican que el ESD exija mayor intervención sobre los sistemas de IA. El CHS garantiza que la decisión definitiva no sea meramente simbólica ni una simple reproducción del resultado algorítmico, sino que exista una instancia humana capaz de revisar, corregir, contextualizar y, si es necesario, apartarse del modelo. De este modo se mantiene la responsabilidad final en el agente humano y se preserva la centralidad de la persona.

4.1.5. Conclusiones del marco teórico: síntesis de hallazgos, mecanismos de reproducción y mitigación de desigualdades

4.1.5.1 Hallazgos centrales y Convergencias teóricas

A partir del recorrido realizado, el marco teórico permite extraer al menos cuatro conclusiones centrales que orientan la tesis:

1. **La naturaleza predictiva de la IA y sus implicaciones jurídicas.** La IA se define, en gran medida, como una tecnología de la predicción, capaz de completar información faltante a partir de datos existentes. Esta funcionalidad la hace especialmente atractiva en el ámbito penal (estimación de riesgo de reincidencia) y judicial (clasificación o *triaje* de casos). Sin embargo, la lógica predictiva entra en

tensión con la exigencia jurídica de justificación normativa de las decisiones públicas, que no pueden reducirse a meras predicciones estadísticas ni a correlaciones opacas.

2. **Riesgo estructural de reproducción de desigualdades.** La literatura muestra de forma consistente que la IA —en particular el *machine learning* y el *deep learning*— corre el riesgo de reproducir y amplificar desigualdades históricas cuando los datos de entrenamiento están sesgados o cuando el pasado se convierte en una "receta para la discriminación algorítmica". En el contexto punitivo colombiano, sistemas como PRISMA, aunque buscan objetividad al centrarse en factores estáticos, exigen un análisis riguroso para descartar fenómenos de discriminación indirecta mediante el uso de *proxies*.

3. **El poder algorítmico y la necesidad de control.** Las teorías de la sociología del castigo, la gubernamentalidad algorítmica y el capitalismo de la vigilancia convergen en caracterizar a la IA estatal como un nuevo dispositivo de poder y control. Este poder no es neutral ni meramente técnico, sino que organiza formas específicas de vigilancia, clasificación y gestión de poblaciones. De ahí la necesidad de articular mecanismos de limitación como la transparencia, la explicabilidad y, especialmente, el Control Humano Significativo (CHS), que permiten enfrentar el problema de la "caja negra" y mantener la responsabilidad en agentes humanos identificables.

4. **Reconfiguración de garantías constitucionales y emergencia de nuevos derechos.** El impacto de la IA en la esfera estatal obliga a repensar categorías clásicas de derechos humanos y fundamentales, en particular la dignidad humana, la igualdad material, la prohibición de discriminación y la protección frente a la manipulación. En este marco se ubican debates como el reconocimiento de neuroderechos y la formulación

del control humano como un derecho instrumental, orientado a garantizar que la automatización no erosione el núcleo del Estado Social de Derecho.

Estas conclusiones preparan el terreno para una reflexión más específica sobre los mecanismos mediante los cuales la IA reproduce o mitiga desigualdades en el ESDD, lo cual se desarrolla a continuación y servirá como base para el análisis empírico posterior.

4.1.5.2 Mecanismos de reproducción y mitigación de desigualdades en el Estado Social de Derecho

A. Reproducción de desigualdades: diseño, datos y la trampa del formalismo

A partir de lo anterior puede verse con mayor claridad que la IA no es una solución neutral; más bien, constituye un espacio donde se cruzan disputas éticas y políticas (Duque, 2025). Por eso resulta útil distinguir, dentro del mismo Estado Social de Derecho, entre mecanismos que reproducen desigualdades y mecanismos que las mitigan.

- **Sesgos en los datos.** Los sistemas de IA están entrenados para replicar patrones de toma de decisiones aprendidos de los datos que alimentan dichos sistemas. Si los datos de entrada contienen sesgos, contradicen estándares de derechos humanos o reflejan prejuicios humanos existentes, perpetúan dichos prejuicios y reproducen condiciones de desigualdad. Otro caso paradigmático, además de los ya vistos, es el de Sisbén IV en Colombia que ilustra cómo los algoritmos pueden profundizar asimetrías históricas cuando se alimentan de datos sesgados (Duque, 2025).
- **Diseño opaco del modelo y trampa del formalismo.** La opacidad y complejidad de algunos sistemas de IA dificultan la identificación de sesgos. Cuando la equidad se reduce a métricas abstractas sin considerar el contexto social y político —la

llamada “trampa del formalismo”— se neutraliza la complejidad y se reproduce la discriminación (Coddou et al., 2025).

- **Instrumentalización de la dignidad.** La instrumentalización de la persona para alcanzar fines privados o lucrativos a costa de los derechos fundamentales —por ejemplo, el uso de datos con sesgos discriminatorios para condicionar el acceso a derechos como educación o salud— reproduce exclusiones y vulnera el mandato de tratar a las personas como fines en sí mismas.

B. Mitigación de desigualdades: gobernanza y control humano

Ahora bien, los mecanismos para mitigar las desigualdades estructurales y proteger las garantías propias del ESDD se centran en el fortalecimiento de la gobernanza y de la presencia humana:

- **Gobernanza participativa y contextualizada.** Una gobernanza participativa, basada en la soberanía de los datos y la justicia cognitiva, puede reorientar el uso de la IA hacia el bien común. Esto incluye integrar explícitamente dimensiones como la distribución de sesgos y el impacto sobre grupos vulnerables y marginados, para evitar la trampa del formalismo y evaluar de forma contextual las métricas de equidad.
- **Control Humano Significativo (CHS).** El CHS actúa como una herramienta jurídica e institucional para lograr la protección real y efectiva de los derechos fundamentales en escenarios de IA. Asegurar la veeduría y la intervención humana desde el diseño hasta la revisión ex post es clave para garantizar la defensa de los derechos, la no discriminación y la justicia, manteniendo la atribución de responsabilidad y la posibilidad de impugnación.

- **Estándares de diseño ético.** El punto de partida de la IA debe ser garantizar que el diseño y la implementación de los sistemas sean compatibles con los derechos humanos, prohibiendo prácticas violatorias desde las fases de diseño y aprendizaje. Se debe asegurar la transparencia, la imparcialidad y la equidad en los métodos operativos, incorporando auditorías, evaluaciones de impacto y mecanismos de corrección de sesgos.

4.2 Antecedentes investigativos

Los antecedentes de investigación se organizan por proximidad temática, comenzando con los estudios que abordan la implementación de la Inteligencia Artificial (IA) en el sistema judicial colombiano, en cuanto dispositivos de poder y control, y continuando con las investigaciones que articulan la discriminación algorítmica y el control humano con las garantías fundamentales del Derecho Constitucional.

A. Herramientas de la Inteligencia Artificial dentro del Sistema Judicial Colombiano: Estudio del caso Pretoria y PRiSMA

Este trabajo de grado, realizado por Torres, Silva y Gómez (2022), tiene como objetivo central analizar la efectividad del uso de herramientas de IA, específicamente PretorIA (en la Corte Constitucional) y PRiSMA (en la Fiscalía General de la Nación). La investigación emplea una metodología de observación descriptiva y explicativa, con un enfoque mixto (cualitativo y cuantitativo), basándose en la recolección y análisis de datos de las entidades judiciales (Torres, et al., 2022). El ámbito de estudio es la funcionalidad jurídico-administrativa de la IA en el contexto judicial colombiano, buscando soluciones a la congestión del sistema.

El aporte principal de este antecedente radica en identificar la necesidad urgente de establecer un código ético o moral para las nuevas tecnologías implementadas en Colombia, dado que la IA aún no está regulada por ley. Además, detalla el funcionamiento de PRISMA, cuyo objetivo es objetivar la solicitud de medidas de aseguramiento y garantizar el principio de igualdad, concentrando las medidas intramurales en individuos con alto riesgo de reincidencia. No obstante, el estudio señala riesgos, como la opacidad del algoritmo de PRISMA y la posibilidad de que conduzca a una discriminación indirecta (Torres, et al., 2022).

Este análisis se articula directamente con la presente investigación, ya que PRISMA constituye un caso relevante a la hora de estudiar la IA estatal en dispositivos de poder, vigilancia y control en Colombia. El mencionado antecedente coincide en la alerta sobre la vulneración del principio de igualdad por sesgos algorítmicos. Por el contrario, a diferencia de su enfoque centrado en la efectividad administrativa y la recomendación ética general, mi estudio indaga específicamente en cómo estas configuraciones técnico-institucionales reconfiguran el contenido de las garantías del Estado Social de Derecho de igualdad material, no discriminación y dignidad humana.

En síntesis**,** de este antecedente tomo el marco de análisis comparado de PRISMA y PretorIA como dispositivos de IA estatal. Sin embargo, persiste la diferencia respecto a la reconstrucción de los umbrales de validez constitucional de la igualdad material, no discriminación y dignidad humana frente a la clasificación automatizada en la justicia penal.

B. ¿Efectivización de los cupos carcelarios?: Aproximación al Sistema Prisma de la Fiscalía General de la Nación

En conexión con el antecedente anterior, el trabajo de Moreno (2019) se concentra en el Sistema PRISMA, una herramienta de *Machine Learning* diseñada para predecir el riesgo de reincidencia criminal y apoyar la imposición de medidas de aseguramiento en centros carcelarios. El objetivo es analizar sus fortalezas y limitaciones en el sistema penal acusatorio colombiano a través de investigación documental y análisis de teorías criminológicas (prevención del delito e incapacitación) y experiencias internacionales tales como COMPAS en Estados Unidos.

El aporte principal del estudio es la descripción crítica de PRISMA, destacando la fortaleza de basarse principalmente en factores de riesgo estáticos y objetivos (historial criminal), lo cual, se postula, minimiza el uso de variables sociodemográficas (proxies) que generan discriminación indirecta. Sin embargo, señala como debilidades cruciales la falta de transparencia sobre el algoritmo, la potencial vulneración del principio de igualdad si el Fiscal utiliza la herramienta de forma selectiva, y la omisión de factores externos y psicológicos en la medición. Por ello, concluye que debe buscarse la democratización de los algoritmos y la transparencia.

En este sentido, el trabajo de Moreno informa mi investigación al proveer una evaluación detallada de PRISMA, un caso central de mi análisis de dispositivos de poder y control punitivo. La documentación sobre los factores de riesgo objetivos versus sociodemográficos se alinea con la pregunta de mi tesis sobre qué rasgos de diseño operan como *proxies* de categorías protegidas. A diferencia de este estudio, que se mantiene en un análisis de criminología y políticas públicas, mi enfoque busca traducir esos riesgos de

opacidad y selección en criterios jurídicos verificables que permitan determinar la validez constitucional de la herramienta.

En conclusión, tomo la distinción entre factores estáticos y objetivos y el riesgo de sesgos por variables sociodemográficas. Sin embargo, subsiste una brecha importante sobre la falta de reconstrucción explícita de los estándares constitucionales para determinar cuándo la clasificación automatizada por riesgo se vuelve cosificante o estigmatizante, lesionando la dignidad humana.

C. Análisis Legal y Jurisprudencial del Impacto de la Discriminación Algorítmica en el Derecho a la Igualdad en la Era de la Inteligencia Artificial

El trabajo de grado de Rodas Florián (2024) realiza un análisis integral sobre el fenómeno de la discriminación algorítmica y sus repercusiones en el derecho a la igualdad, tanto a nivel internacional como en la jurisprudencia de la Corte Constitucional colombiana. Su metodología es cualitativa y se basa en el análisis legal y la evaluación de decisiones judiciales y casos emblemáticos. El ámbito de estudio abarca la legislación y jurisprudencia que protegen el derecho a la igualdad en la era de la IA (Rodas Florián, 2024).

El aporte principal reside en establecer el marco conceptual que distingue entre igualdad formal, igualdad material y la prohibición de discriminación en la Constitución colombiana. Además, define la discriminación algorítmica como la toma de decisiones automatizadas que perpetúa sesgos. Documenta casos relevantes como COMPAS, Amazon y Google, utilizándolos para identificar lagunas legales y la necesidad de salvaguardar los derechos fundamentales (Rodas Florián, 2024).

Esta investigación coincide con mi tesis en el reconocimiento de la discriminación algorítmica como una amenaza al derecho a la igualdad. Las categorías de igualdad material y discriminación indirecta son fundamentales para el análisis constitucional que propongo. Por ello, este antecedente sienta una base teórica sólida. A diferencia de este trabajo, que se enfoca en el análisis del impacto general y la identificación de vacíos, mi enfoque utiliza estas categorías para evaluar cómo la IA, al ser incorporada en dispositivos de vigilancia y control estatales (que implican ejercicio de poder punitivo), transforma o erosiona los estándares constitucionales sustantivos.

En concreto, de este precedente tomo las categorías jurisprudenciales colombianas de discriminación indirecta e igualdad material como lentes para el análisis de los sesgos en la IA estatal. La diferencia es que se requiere un enfoque centrado en la reconstrucción crítica de los umbrales de validez jurídica para determinar cuándo la clasificación automatizada por riesgo se vuelve cosificante o estigmatizante, lesionando la dignidad humana.

D. El derecho al control humano en la inteligencia artificial

Esta investigación aborda el problema jurídico de reconocer y regular el control humano en la inteligencia artificial como un nuevo derecho en el ordenamiento colombiano. El objetivo es formular una propuesta de regulación (*lege ferenda*) que garantice la veeduría humana en el diseño y uso de la IA, previniendo así la afectación a derechos ya existentes. Se utiliza una metodología que combina el rastreo de doctrina internacional con el análisis de compatibilidad con el marco jurídico colombiano (Sánchez 2021).

El aporte central es la conceptualización del control humano como una garantía de veeduría y supervisión, justificada en su relación instrumental con la dignidad humana y la necesidad de proteger la igualdad y el debido proceso ante la opacidad algorítmica. Se identifica que el control humano debe tener funciones preventivas (monitoreo en el diseño) y correctivas (revisión de decisiones automatizadas). Asimismo, se alerta sobre cómo la IA, utilizada en la justicia (Prometea y PRISMA), puede amenazar el debido proceso y reflejar prejuicios humanos (Sánchez 2021).

Este trabajo se relaciona con mi tesis al enfatizar que la intervención humana es crucial para la protección de la igualdad y la dignidad, lo cual es un componente esencial de las salvaguardas que mi investigación busca delimitar. Por el contrario, a diferencia de este antecedente, que se enfoca en la propuesta de creación de un nuevo derecho fundamental (*lege ferenda*), mi análisis se centra en determinar qué nivel de control humano significativo es una condición de legitimidad para las decisiones públicas asistidas por IA dentro del marco constitucional vigente.

Por lo tanto, la investigación realizada por Sánchez Vásquez (2021). representa un importante antecedente a la presente investigación, toda vez que lleva a cabo la conceptualización del control humano significativo y la distinción entre sus funciones preventiva y correctiva, usándolas como criterios de evaluación de la gobernanza de la IA estatal.

E. Inteligencia Artificial, Ética y Regulación Jurídica: Una Mirada Desde el Derecho Constitucional Colombiano

Barco, Mendoza y Urbano (2023) abordan la necesidad de establecer un marco normativo para la IA en Colombia, analizando los lineamientos internacionales y su impacto en la vulneración de derechos fundamentales. Utilizan una metodología socio-jurídica e inductivo-deductiva, examinando el avance de la IA en el país (ej. uso en el sector judicial) en relación con los principios constitucionales.

El trabajo concluye que, si bien Colombia ha adherido a principios éticos internacionales (OCDE por medio del CONPES 3975), existe un vacío normativo serio que ha llevado a la utilización de la IA por analogía, lo que acrecienta la vulneración de derechos como la no discriminación y la dignidad humana. El principal aporte es la identificación de los retos normativos, enfatizando el riesgo de sesgo y discriminación en tecnologías como el Reconocimiento Facial (TFR). Propone que la futura regulación debe fundarse en principios constitucionales de dignidad, transparencia y justicia social (Barco 2023).

Este antecedente subraya la relevancia de mi investigación al constatar que la IA compromete activamente los derechos de no discriminación y la dignidad humana en el contexto colombiano. Coincide en que la IA en la seguridad pública (TFR) genera riesgos graves (Barco 2023). A diferencia de este estudio, que busca proponer las características de la ley (*marco normativo*), mi tesis se centra en el análisis crítico de las configuraciones de poder, vigilancia y control ya existentes, para reconstruir los estándares constitucionales que deben limitar su uso,

En particular, tomo la alerta sobre la falta de regulación específica en Colombia y la identificación de la vulneración de la dignidad como resultado de la discriminación algorítmica en tecnologías de seguridad. La brecha que mi estudio aborda es la necesidad de

examinar la reconfiguración de las garantías constitucionales más allá de la mera constatación de la vulneración, para establecer criterios de lectura jurídica que permitan valorar la conformidad de las prácticas estatales con los preceptos fundamentales.

F. Tensiones y realidades sobre la vulneración de los derechos fundamentales a falta de regulación de la inteligencia artificial (IA) en Colombia

El artículo de Aponte Fonseca (2023) identifica las tensiones y realidades sobre la vulneración de derechos fundamentales ante la falta de regulación de la IA en Colombia. El enfoque es descriptivo y analítico, examinando contextos internacionales de debate sobre la regulación de la IA para luego analizar su implementación en el sistema jurídico colombiano.

El principal aporte es la contextualización de la implementación de la IA en las altas cortes colombianas (mencionando a Prometea/Pretoria y PRISMA), como apoyo a la descongestión judicial. El estudio concluye que, aunque la IA puede servir a la eficiencia, su uso inadecuado sin regulación pone en riesgo derechos fundamentales (intimidad, igualdad, debido proceso, dignidad). En este sentido, subraya que la IA debe ser una herramienta de apoyo al juez, cuya labor no puede ser reemplazada (Fonseca 2023).

Este antecedente respalda mi problema de investigación al confirmar las tensiones generadas por la implementación de sistemas de IA estatales (PRISMA, PretorIA) ante la ausencia de regulación que proteja los derechos fundamentales. La identificación de la IA como fuente de discriminación inadvertida y sesgos en la toma de decisiones se alinea con el núcleo de mi tesis sobre la no discriminación y la igualdad material. Por consiguiente, el

trabajo de Aponte informa mi enfoque al mapear la afectación a la dignidad humana por la intromisión en la esfera privada y la relación de poder generada por la IA.

4.3 Antecedentes Normativos y Jurisprudenciales

A continuación, se sistematizan las fuentes que regulan de forma expresa la Inteligencia Artificial (IA) y que, por su estructura y contenido, sirven de referencia directa para evaluar garantías como igualdad y no discriminación, transparencia, explicabilidad y control humano significativo en el diseño y uso de sistemas algorítmicos.

En el plano comparado, el Reglamento 2024/1689 de la Unión Europea, conocido como Ley de Inteligencia Artificial, se ha convertido en el referente normativo más completo. Este instrumento adopta un enfoque basado en el nivel de riesgo y se estructura en cuatro grandes ejes: (i) prohibiciones, como los sistemas de puntuación social, que asignan una calificación global a las personas para concederles o negarles beneficios, y la manipulación de personas vulnerables; (ii) obligaciones reforzadas para los sistemas de alto riesgo, relativas a la gestión de riesgos, la gobernanza y calidad de los datos, la documentación, la supervisión humana y la exigencia de exactitud y robustez; (iii) deberes de transparencia en la interacción entre personas y sistemas de IA y en el uso de contenidos sintéticos (textos, imágenes o videos generados artificialmente); y (iv) reglas específicas para los modelos de IA de propósito general, esto es, sistemas reutilizables en múltiples ámbitos, incluidos aquellos considerados de riesgo sistémico.

Desde el punto de vista normativo, resultan especialmente relevantes el artículo 5, que establece las prácticas prohibidas; los artículos 8 a 15, que fijan las obligaciones para los

sistemas de alto riesgo en conexión con el artículo 6 y el Anexo III; el artículo 50, relativo a la transparencia; y el Capítulo V, dedicado a los modelos de inteligencia artificial de propósito general. En conjunto, este diseño normativo funciona como una plantilla aplicable a los usos estatales de la IA que pueden afectar la igualdad material, la no discriminación y la dignidad, pues exige que estos sistemas sean explicables, trazables y sometidos a un control humano significativo tanto antes como después de su puesta en funcionamiento.

De forma complementaria, el Convenio Marco del Consejo de Europa sobre IA, Derechos Humanos, Democracia y Estado de Derecho (CETS 225, 2024) fija por primera vez en un tratado vinculante la obligación de asegurar, durante todo el ciclo de vida de los sistemas de IA, la evaluación de riesgos/impactos, la transparencia y supervisión, y la disponibilidad de recursos efectivos cuando la IA afecte derechos. Este umbral convencional funciona como parámetro de legitimidad para decisiones automatizadas que incidan en grupos históricamente vulnerados.

En el nivel intergubernamental, la Recomendación de la UNESCO sobre la Ética de la IA (2021) y la Recomendación de la OCDE sobre IA (2019) consolidan un haz de principios hoy internalizados por legislaciones y políticas: no discriminación e inclusión, transparencia/explicabilidad, seguridad/robustez y rendición de cuentas, junto con herramientas de implementación (p. ej., evaluaciones de impacto y gobernanza del ciclo de vida). Su valor para la tesis es metodológico: traducen los derechos en criterios operativos para auditar sesgos y barreras de impugnación en sistemas sociotécnicos.

En la órbita de cooperación y estándares técnicos, el Proceso de Hiroshima del G7 (Principios y Código de Conducta, 2023) impulsa para la IA avanzada y generativa

exigencias de prueba, documentación, reporte de capacidades y limitaciones, gestión de riesgos y comunicación de vulnerabilidades e incidentes. Aunque no tiene la fuerza vinculante de una ley, sí cumple una función relevante como estándar técnico-operativo, pues traduce preocupaciones sobre seguridad, sesgos, transparencia y rendición de cuentas en pautas verificables para el diseño, despliegue y supervisión de estos sistemas.

En el mismo sentido, como telón multilateral, la Resolución 78/265 (2024) de la Asamblea General de la ONU insta a aprovechar una IA segura, fiable y confiable para el desarrollo, a abstenerse de usos incompatibles con derechos humanos y con el DIH, y a reforzar privacidad, monitoreo de riesgos, cooperación y creación de capacidades. Su valor para la tesis es anclar el deber estatal de debida diligencia y la necesidad de marcos interoperables cuando los sistemas se emplean en funciones públicas sensibles.

Finalmente, la AGNU 78/265 (2024) sitúa la IA en una agenda común de seguridad, confiabilidad y derechos humanos, llamando a evitar sistemas incompatibles con el DIDH y a robustecer la cooperación y el monitoreo de riesgos. Para una tesis centrada en dispositivos estatales de poder, esta resolución refuerza el deber de debida diligencia y la exigencia de marcos de transparencia y rendición de cuentas.

Por su parte, en Colombia, existen instrumentos expresos que aterrizan estos estándares. La Circular Externa SIC 002 de 2024 fija lineamientos específicos para el tratamiento de datos personales en sistemas de IA e introduce un test de cuatro pasos — idoneidad, necesidad, razonabilidad y proporcionalidad— como filtro de legitimidad del tratamiento, además de exigir evaluaciones de impacto en privacidad cuando el riesgo lo demande, medidas de privacidad desde el diseño y deberes reforzados de información y

transparencia frente a los titulares. Con ello, este instrumento opera como parámetro para exigir justificación ex ante, mitigación de sesgos y trazabilidad de decisiones automatizadas con efectos diferenciados.

En la Rama Judicial, el Acuerdo PCSJA24-12243 del 16 de diciembre de 2024 adopta lineamientos para el uso responsable y ético de IA declara que la IA no reemplaza la función jurisdiccional ni la valoración probatoria; impone control humano, seguridad y protección de datos; y ordena un registro de iniciativas (art. 14) para trazabilidad pública, lo cual para un enfoque garantista, aporta criterios institucionales concretos de explicabilidad y supervisión en la toma de decisiones judicial asistida por IA.

Como marco de política pública, el CONPES 3975 de 2019, que establece la Política Nacional para la Transformación Digital e Inteligencia Artificial, resulta clave al definir la IA como un campo de la informática dedicado a resolver problemas cognitivos asociados a la inteligencia humana, basado en el desarrollo de sistemas, datos y algoritmos; Con posterioridad, el CONPES 4144 de 14/02/2025 fija la Política Nacional de IA, alineando capacidades, ética y gestión de riesgos en el sector público (sin ser ley formal, condiciona programas y presupuestos, y estandariza expectativas de gobernanza y transparencia en proyectos de IA).

En ambos documentos de política pública, se propende por desarrollar condiciones habilitantes para preparar a Colombia para los cambios económicos y sociales que conlleva la IA e impulsar otras tecnologías de la Cuarta Revolución Industrial (4RI). Para este fin, solicitó el diseño de un marco ético transversal que guíe el diseño, desarrollo, implementación y evaluación de sistemas de IA, siguiendo los principios de la OCDE. Dicho

marco debe contemplar, como mínimo, la justicia, la transparencia, explicabilidad, la libertad, la responsabilidad, la inclusión y el rol de los derechos humanos, buscando un balance entre la protección de los ciudadanos y el fomento de la innovación.

En el plano legislativo interno, resulta especialmente relevante el Proyecto de Ley 043 de 2025 (Congreso de la República de Colombia, 2025) "*Por medio de la cual se regula la Inteligencia Artificial en Colombia para garantizar su desarrollo ético, responsable, competitivo e innovador, y se dictan otras disposiciones*". Esta iniciativa propone un marco jurídico integral y adaptativo para el desarrollo, implementación y uso de la IA en Colombia, articulando criterios técnicos con la protección de derechos fundamentales y el desarrollo productivo. Entre sus ejes se encuentran la creación de una Autoridad Nacional para la IA liderada por MinCiencias, una clasificación de sistemas en cuatro niveles de riesgo (crítico, alto, limitado y bajo), la habilitación parametros regulatorios, la incorporación transversal de la IA en el sistema educativo, medidas de protección del trabajo y reconversión laboral, así como principios de transparencia, explicabilidad, gobernanza responsable, no discriminación e inclusión con enfoque diferencial. Aun en trámite legislativo, el proyecto constituye un referente normativo clave para anticipar la dirección de la futura regulación de la IA en Colombia y contrastar su coherencia con los estándares internacionales y de política pública previamente descritos.

Y, para la transparencia algorítmica en el Estado, la Directiva Conjunta 007 del 30 de septiembre de 2025 de la Procuraduría General de la Nación y la Defensoría del Pueblo, establece estándares mínimos de publicación sobre lógicas generales, datos y garantías, y prevé análisis de impacto algorítmico; su objetivo es asegurar explicabilidad y acceso

suficiente para el control ciudadano, en sintonía con el derecho de acceso a la información pública.

De otro lado, *La Guía con lineamientos generales para el uso de tecnologías emergentes* del Ministerio de Tecnologías de la Información y las Comunicaciones (MinTic) opera como un instrumento de *soft law* diseñado para orientar a las entidades públicas (nacionales y territoriales) en la adopción y uso de nuevas herramientas digitales. Esta guía establece un camino de implementación estructurado en fases (Comprender, Diseñar, Habilitar e Implementar) y enfatiza la necesidad de alinear la estrategia de adopción con los propósitos de la política de gobierno digital. Su objetivo es lograr servicios más eficientes, intuitivos y seguros, que tomen decisiones basadas en datos y generen valor público.

En conjunto, estos instrumentos conforman un piso normativo que nombra y regula la IA de forma expresa y que, de forma articulada permiten: (i) exigir proporcionalidad material, (ii) verificar controles ex ante y supervisión humana, (iii) demandar transparencia y (iv) anclar la evaluación de sesgos y afectaciones sobre igualdad, no discriminación y dignidad a un arreglo institucional verificable.

Por último, se destacan las siguientes fuentes jurisprudenciales nacionales en la materia: Sentencia STC17832-2025 de la Sala de Casación Civil, Agraria y Rural de la Corte Suprema de Justicia y las Sentencias T-323/24 y T-067/25 de la Corte Constitucional.

La primera de ellas, es decir, la Sentencia T-323/24 es fundamental al establecer criterios orientadores para el uso adecuado de herramientas de Inteligencia Artificial (IA),

como ChatGPT, por parte de los despachos judiciales en Colombia, enfocándose en la protección del derecho fundamental al debido proceso.

La relevancia de esta decisión radica en que ordena la aplicación de doce principios esenciales, incluyendo la Transparencia (obligación de evidenciar el uso y alcance de la IA), la Responsabilidad (exigiendo que el usuario de la IA esté capacitado, verifique la información y dé cuenta de su origen e idoneidad), la Privacidad (custodiando datos personales y sensibles), y el principio de No sustitución de la racionalidad humana (prohibiendo que la IA reemplace las labores jurisdiccionales indelegables del juez).

Además, la Sentencia exhorta a los jueces a evaluar el uso apropiado de estas herramientas y ordena al Consejo Superior de la Judicatura divulgar una guía o lineamiento acorde con estos aspectos para la implementación de la IA generativa en la Rama Judicial.

Por su parte la sentencia T-067/25 marca un hito en la regulación de la tecnología estatal al garantizar el derecho de acceso a la información pública respecto al código fuente de aplicaciones informáticas gubernamentales, como CoronApp. Este fallo establece la transparencia algorítmica como una garantía fundamental para asegurar el uso adecuado y razonable de los datos personales y prevenir decisiones arbitrarias o discriminatorias por parte de los Sistemas Automatizados de Toma de Decisiones (SDA) utilizados por entidades públicas.

La Corte determinó que, al negar el acceso al código fuente, las autoridades afectaron el control ciudadano y la capacidad de las personas para verificar la precisión, seguridad y uso correcto de sus datos. El fallo enfatiza que la transparencia algorítmica (activa y pasiva)

debe regirse por un principio de divulgación máxima cuando se trata de SDA a cargo del Estado, permitiendo el examen del desempeño de la herramienta tecnológica utilizada. Consecuentemente, ordenó la expedición de lineamientos sobre transparencia algorítmica en sistemas utilizados por el Estado, a cargo de la Agencia Nacional Digital y el Ministerio Público.

Finalmente, la Sentencia STC17832-2025 de la Sala de Casación Civil, Agraria y Rural de la Corte Suprema de Justicia consolida una línea garantista sobre el deber de motivación y la correcta utilización de fuentes jurisprudenciales en las decisiones judiciales. En este fallo, la Corte ampara a la accionante al constatar que el Tribunal Superior de Sincelejo construyó la terminación del proceso por desistimiento tácito con base en citas inexistentes de precedentes (STC13560-2023 y STC4734-2025), configurando un defecto de motivación y una vía de hecho. La decisión enfatiza que los jueces deben verificar con alta diligencia la veracidad y fidelidad de las fuentes que emplean —incluida la jurisprudencia y, cuando se utilicen, las salidas de sistemas de IA— retomando expresamente la Sentencia T-323 de 2024 sobre el riesgo de "alucinaciones" y la necesidad de control humano en el uso de estas herramientas. De este modo, la STC17832-2025 vincula el principio de debida motivación con la confianza legítima en la administración de justicia y refuerza la idea de que el uso de tecnologías como la IA no exime al funcionario de su responsabilidad en la verificación, selección y justificación de los fundamentos que soportan sus providencias.

5. METODOLOGÍA

5.1 Naturaleza y enfoque de la investigación

La presente tesis se desarrolla bajo un enfoque jurídico-hermenéutico, de carácter cualitativo (Sampieri & Mendoza, 2018; Abásolo, 2023), lo cual se traduce en dos ejes fundamentales. El primero de ellos significa que es una investigación jurídica (Clavijo Cáceres et al., 2014) en cuanto analiza categorías normativas y garantistas propias del Estado Social de Derecho (igualdad material, no discriminación y dignidad humana), y verifica en qué medida dichas garantías se ven preservadas, erosionadas o transformadas cuando el Estado incorpora sistemas de inteligencia artificial en sus dispositivos de poder, vigilancia y control.

Y el segundo de ellos se circunscribe a una investigación hermenéutica-documental (Clavijo Cáceres et al., 2014), en tanto trabaja sobre interpretación y reconstrucción crítica de fuentes jurídicas (Constitución, bloque de constitucionalidad, jurisprudencia, normas, lineamientos institucionales), técnicas (descripciones de modelos algorítmicos, evaluaciones de impacto, documentos de gobernanza tecnológica) y doctrinales (teorías sobre poder, vigilancia, control, dominación, biopolítica, gubernamentalidad, y doctrina jurídica sobre la IA).

Este enfoque hermenéutico no se limita a describir textos; busca hacer explícito el sentido constitucional de los efectos de la IA: cuándo esa tecnología opera como herramienta de gestión estatal legítima y cuándo activa lógicas de disciplinamiento, estigmatización o exclusión incompatibles con los principios garantistas del Estado Social de Derecho. Este proceso de análisis e interpretación busca **"***descubrir los conceptos, categorías, temas y patrones presentes en los datos, así como sus vínculos, a fin de otorgarles sentido, interpretarlos y explicarlos en función del planteamiento del problema***"** (Sampieri & Mendoza, 2018, p. 465).

Al mismo tiempo, la tesis incorpora un componente empírico de estudio de casos, entendido no como medición estadística exhaustiva del fenómeno estatal, sino como análisis jurídico-constitucional situado sobre configuraciones técnico-institucionales concretas, tanto en Colombia como en otras jurisdicciones comparables. Y es que la elección de este enfoque permite abordar la facticidad del estudio científico en el Derecho, es decir, "*los hechos que originan las normas que se presentan y generan consecuencias en contextos históricos y sociales específicos*" (Clavijo Cáceres et al., 2014, p. 43), en tanto analiza casos que ya han generado controversia por sesgo, opacidad o impacto desigual sobre poblaciones vulnerables.

5.2 Diseño metodológico y articulación con los objetivos

Así las cosas, el desarrollo de la tesis se estructura en cuatro capítulos correspondientes respectivamente a cada uno de los objetivos específicos formulados.

Capitulo I. Analizar la incorporación de sistemas de inteligencia artificial en dispositivos estatales de poder, vigilancia y control, en el marco de la sociología de del control y del castigo (Objetivo específico 1).

En esta capitulo se reconstruye y sistematiza el marco teórico y conceptual necesario para leer críticamente la relación entre inteligencia artificial y poder estatal. Se abordan, entre otras, las nociones de poder, dominación, vigilancia, control, biopolítica y gubernamentalidad; así como categorías operativas sobre inteligencia artificial (definiciones de IA, tipologías de sistemas, modos de intervención estatal algorítmica). El propósito de esta fase no es meramente descriptivo, sino analítico: fijar las categorías de lectura que luego serán aplicadas a los casos empíricos y a la evaluación constitucional. Esto da el andamiaje

interpretativo para comprender por qué la IA no es un recurso técnicamente neutro, sino un dispositivo que puede reproducir desigualdades históricas bajo la apariencia de eficiencia.

Capítulo II. Caracterización técnico-institucional (Objetivo específico 2)

En este capítulo se identifica y caracteriza las configuraciones técnico-institucionales de los sistemas de IA seleccionados. Se examina, para cada caso: la procedencia y el tratamiento de los datos; el diseño del modelo (cómo puntúa, perfila o clasifica); los esquemas de gobernanza institucional, es decir, quién lo opera, con qué controles y bajo qué marco jurídico declarado.

En paralelo se identifican y registran con base en la evidencia documental los mecanismos de afectación (por ejemplo, sesgos de entrenamiento, estigmatización de ciertos grupos) y de mecanismos de contención documentados (por ejemplo, evaluaciones de impacto en derechos, trazabilidad de decisiones, transparencia operativa, auditorías) sin emitir todavía la valoración garantista definitiva, que se reserva para el capítulo IV.

Capítulo III. Examinar los hallazgos entre casos nacionales e internacionales para identificar las condiciones para la protección de las garantías del ESDD (Objetivo específico 3)

Este capítulo desarrolla el examen de los casos nacionales e internacionales seleccionados, entendido como un contraste de patrones y condiciones a partir de los hallazgos consignados en el Capítulo II. El propósito es describir e identificar la forma en la cual las condiciones técnico-institucionales de la IA estatal terminan preservando, erosionando o transformando las garantías del Estado Social de Derecho. El análisis atiende especialmente a si el sistema tiende a reproducir o mitigar desigualdades estructurales frente

a poblaciones históricamente vulneradas, punto que es central en el planteamiento del problema.

A partir de esa comparación se reconstruyen umbrales de validez constitucional, es decir, criterios mínimos exigibles para que una herramienta algorítmica estatal sea jurídicamente aceptable desde las garantías de igualdad material, no discriminación y dignidad humana. Estos umbrales se justifican a partir de la construcción dogmática, teórica y jurisprudencial de dichas garantías y de su interacción con el uso estatal de IA realizada en el Capítulo I

Este capítulo es fundamental porque proporciona la base argumentativa para una lectura garantista posterior, no se limita al caso colombiano aislado, sino que sitúa a Colombia en una cartografía de prácticas estatales contemporáneas de vigilancia, control y castigo mediadas por IA.

Capítulo IV. Evaluación garantista caso por caso (Objetivo específico 4). En el último capítulo, la tesis evalúa en detalle cómo las configuraciones técnico-institucionales identificadas inciden sobre las garantías del Estado Social de Derecho —igualdad material, no discriminación y dignidad humana—. Esta evaluación operará sobre cada caso concreto y busca dos resultados:

1. Precisar cuáles estándares constitucionales están comprometidos en su aplicación real.

2. Determinar el nivel de cumplimiento y las desviaciones relevantes frente a dichos estándares.

En términos metodológicos, aquí se da cierre al trabajo: se aplica el marco conceptual (Capítulo I) a la descripción empírica (Capítulo II), se contrasta con las condiciones reconstruidas (Capítulo III), y se formula una lectura jurídico-garantista final. Esta fase responderá de manera directa a la pregunta de investigación formulada en el planteamiento del problema sobre cómo la IA reconfigura las garantías del Estado Social de Derecho.

5.3 Criterios para la selección de los casos

Los casos se seleccionaron con base en los siguientes criterios, que emergen del planteamiento del problema:

- **Relevancia garantista:** casos en los que la IA incide (o pretende incidir) en decisiones con efectos materiales sobre derechos y libertades, especialmente frente a poblaciones históricamente vulneradas.

- **Controversia jurídica o social documentada:** existencia de señalamientos en torno a sesgo, discriminación indirecta, opacidad, ausencia de control humano significativo, vigilancia masiva o estigmatización sobre grupos poblacionales determinados.

- **Valor analítico comparado:** disponibilidad de elementos que permitan contrastar Colombia con otras jurisdicciones donde tribunales, entes de control o litigios estratégicos hayan intervenido, por ejemplo, declarando ilegal un sistema, imponiendo estándares de transparencia o cuestionando su proporcionalidad.

Esta estrategia evita una aproximación puramente teórica y una puramente cuantitativa: los casos son usados como “sitios de observación jurídica”, donde se hace

visible cómo operan las tensiones entre eficiencia técnica, control social y garantía constitucional.

5.4 Instrumentos de análisis y matrices de sistematización

Con el fin de asegurar consistencia, comparabilidad y trazabilidad entre los capítulos, la investigación emplea un conjunto de matrices de sistematización. Estas matrices permiten: (i) extraer evidencia de fuentes jurídicas, institucionales y técnicas; (ii) reconstruir, caso por caso, la configuración técnico-institucional de cada sistema; y (iii) formular un juicio garantista sobre la conformidad o disconformidad de cada práctica con las garantías del Estado Social de Derecho.

5.4.1. Matriz de caracterización técnico-institucional de los casos (Capítulo II)

Propósito: describir y reconstruir cómo opera el sistema (datos,modelo,gobernanza) e identificar mecanismos de afectación y contención. Se diligencia una vez por cada caso a manera de síntesis.

Tabla 1.
Matriz de caracterización técnico-institucional de los casos

Dimensión	Preguntas guía	Indicadores observables
Finalidad declarada y contexto	¿Qué problema pretende resolver? ¿En qué decisión/incidencia estatal se inserta?	Objetivo formal y ámbito de aplicación (seguridad, asistencia social, justicia penal).
Población y/o territorio impactado	¿A quién aplica? ¿Se focaliza territorial o poblacionalmente?	Criterios de selección, despliegue por zonas, grupos más expuestos.
Datos: origen y calidad	¿De dónde vienen los datos? ¿Son pertinentes, representativos, actualizados?	Fuentes, minimización, calidad, variables sensibles y proxies.

Dimensión	Preguntas guía	Indicadores observables
Tratamiento de datos	¿Cómo se recolectan, cruzan, conservan y comparten?	Finalidades, retención e interoperabilidad.
Modelo y arquitectura	¿Qué tipo de modelo (reglas/ML/DL)? ¿Qué variables usa?	Tipo de sistema y características.
Salida del sistema	¿Qué produce (score (puntaje), alerta, recomendación)?	Umbrales, categorías y explicación disponible
Rol en la decisión	¿La salida decide, recomienda o prioriza?	Automatización de facto, dependencia institucional
Gobernanza, evaluación de impacto y responsabilidades	¿Quién opera? ¿Quién responde? ¿Qué controles y evaluaciones existen?	Cadena de mando, evaluaciones, auditorías, supervisión y contratación.
Transparencia y trazabilidad	¿Es auditable? ¿Hay explicabilidad y registro?	Logging, acceso a reglas, explicaciones al afectado
Control humano significativo (CHS)	¿Dónde interviene el humano (diseño-uso-revisión)?	Supervisión ex ante, operación, revisión ex post, posibilidad de apartarse de la decisión.
Evidencia de impactos	¿Hay evidencia de sesgos/errores/afectaciones?	Falsos positivos, disparidades, suspensiones, litigios.

5.4.2. Matriz comparada para reconstrucción de umbrales de validez constitucional (Capítulo III)

Propósito: convertir hallazgos intercasos en condiciones mínimas exigibles (umbrales) para que el uso estatal de IA sea compatible con igualdad material, no discriminación y dignidad.

Tabla 2
Matriz comparada para reconstrucción de umbrales de validez constitucional

Garantía	Umbral (condición mínima)	Justificación constitucional (criterio)	Señales de incumplimiento
Igualdad material			
No discriminación			
Dignidad humana			
Control humano significativo			

5.4.3. Matriz de evaluación garantista de los casos (Capítulo IV)

Propósito: aplicar los umbrales reconstruidos y producir un juicio final por caso: cumple / cumple parcialmente / no cumple, con desviaciones y efectos sobre garantías.

Tabla 3
Matriz de evaluación garantista de los casos

Caso	Evidencia relevante	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
SyRI				
AFR Locate				
COMPAS				
PRiSMA				
Fiscal Watson				

5.5 Fuentes de información

La investigación se nutre de fuentes de tres tipos:

1. **Fuentes normativas y jurisprudenciales:** Constitución, bloque de constitucionalidad, jurisprudencia relevante, normas administrativas y documentos

regulatorios sobre vigilancia, seguridad, política criminal, tratamiento de datos, uso institucional de IA. Estas fuentes permiten fijar los estándares de igualdad material, no discriminación y dignidad humana que estructuran el análisis.

2. **Fuentes institucionales y técnicas:** manuales de funcionamiento de sistemas algorítmicos, contratos o convenios de adopción tecnológica, evaluaciones de impacto en derechos o datos, lineamientos internos sobre uso de IA, directrices de gobernanza, informes de auditoría, reportes de efectividad. Estas fuentes permitirán reconstruir las configuraciones técnico-institucionales y el grado de control humano significativo reportado oficialmente.

3. **Fuentes doctrinales y de teoría crítica:** literatura sobre poder, vigilancia, control social, biopolítica, dominación y gubernamentalidad tecnológica; literatura jurídica sobre IA y garantías en el Estado Social de Derecho; y análisis críticos sobre gobernabilidad algorítmica, neutralidad aparente y perfilamiento de riesgo. Estas fuentes permiten conceptualizar la IA como dispositivo de poder y no solo como herramienta técnica.

6. Capítulo II. Caracterización técnico-institucional de los sistemas de inteligencia artificial analizados

El Capitulo II desarrolla el Objetivo específico 2, consistente en identificar y caracterizar las configuraciones técnico-institucionales de los sistemas de IA seleccionados, algunos aún en funcionamiento y otros suspendidos. En esta etapa no se realiza aún la evaluación garantista definitiva; el propósito inmediato es reconstruir con evidencia cómo operó u opera cada sistema en la práctica, qué datos utiliza, qué tipo de intervención

algorítmica realiza (puntuación, clasificación, perfilamiento, priorización o alertas) y bajo qué arreglos institucionales de gobernanza se despliega. Este capítulo constituye, por tanto, el insumo empírico-documental que permite realizar el análisis del Capítulo III y la evaluación caso por caso del Capítulo IV.

La caracterización se desarrolla en dos niveles complementarios: (i) una exposición analítica que se sustenta fuentes primarias y técnicas; y (ii) fichas homogéneas como síntesis estandarizada y cierre por caso. Para garantizar comparabilidad y rigor cada caso se reconstruirá con base en la Matriz del Capítulo II (Tabla 1), que guía la extracción de evidencia sobre finalidad y contexto, población impactada, datos y tratamiento, arquitectura del modelo, salida y rol en la decisión, gobernanza y controles, transparencia y trazabilidad, control humano significativo e impactos documentados.

La reconstrucción de cada caso se apoyará en una estrategia de triangulación documental. Se priorizan decisiones judiciales, normas, documentos institucionales, evaluaciones de impacto, anexos contractuales, auditorías y material técnico verificable; y se complementa con literatura académica y reportes especializados. Cuando exista información no pública o inaccesible, ello se registrará explícitamente como hallazgo relevante en materia de transparencia, trazabilidad y control.

En coherencia con el diseño por capítulos, el conjunto de fichas del Capítulo II permite, en el Capítulo III, identificar patrones y convertirlos en umbrales mínimos de validez. Dichos umbrales se aplicarán posteriormente en Capítulo IV para producir un juicio final por caso el cual se sintetiza en sus respectivas tablas. En este proceso el control humano significativo, como fue establecido en el Capítulo I, se tratará como condición instrumental

que hace exigible y controlable las garantías sustantivas de igualdad material, no discriminación y dignidad humana.

6.1 Caso COMPAS (Estados Unidos)

6.1.1. Finalidad declarada y decisión estatal afectada

COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) es un sistema de evaluación de riesgos y necesidades de cuarta generación diseñado como respuesta técnica ante las demandas de ineficiencia y hacinamiento sistémico de los organismos de justicia penal en Estados Unidos (Equivant, 2017). Fue desarrollado originalmente en 1998 por Northpointe Inc., posteriormente denominada Equivant.

Formalmente, pretende resolver el desafío de organizar y procesar volúmenes masivos de información sobre poblaciones supervisadas para optimizar la gestión de casos y mejorar la seguridad pública mediante la predicción de la reincidencia (Equivant, 2017). Además, se presenta como una solución “científica” orientada a reducir el sesgo humano irracional en la toma de decisiones "ad hoc"[13], las cuales a menudo se basan en la intuición o el instinto en lugar de datos probados, lo que las hace abiertamente inconsistentes (State v. Loomis, 2016). Esta promesa, sin embargo, ha sido discutida críticamente en la literatura,

[13] Las decisiones humanas "*ad hoc*" en el contexto judicial se refieren a determinaciones basadas en la intuición, el instinto y el sentido personal de justicia, en lugar de apoyarse en información estadística completa y precisa. De acuerdo con las fuentes, este tipo de juicios ocurren cuando conclusiones que pueden parecer "intuitivamente correctas en el momento" se toman sin la consideración de ningún tipo de hechos probados, dejando decisiones críticas, al arbitrio de la práctica estándar de un oficial o la corazonada de un juez. Históricamente, el sistema de justicia ha buscado mitigar estas decisiones subjetivas y potencialmente inconsistentes mediante el uso de evaluaciones estadísticas objetivas, bajo la premisa de que estas herramientas organizan la información de manera más consistente y son superiores al juicio humano tradicional o no estructurado.

tanto por sus supuestos sobre neutralidad y sesgo como por sus límites empíricos (Mayson, 2018; Dressel & Farid, 2018; Washington, 2019).

De acuerdo con la Guia para Profesionales de COMPAS Core (2017) de Equivant, el sistema es un instrumento actuarial de cuarta generación[14] destinado a apoyar la evaluación de riesgos y necesidades de personas vinculadas al sistema penal estadounidense. Esta plataforma integrada de gestión de casos procesa información proveniente de entrevistas y registros oficiales para generar escalas de riesgo de reincidencia (general y violenta) y escalas de necesidades criminógenas[15] con el objetivo explícito de mejorar consistencia y eficiencia en la en la administración de justicia penal, sirviendo como soporte en decisiones críticas sobre la ubicación, supervisión y manejo de casos, incluyendo su uso en informes pre-sentenciales.

El sistema se inserta en decisiones estatales fundamentales que distribuyen cargas de vigilancia y beneficios de libertad a través de todo el "continuo" de la justicia penal (Angwin et al., 2016). En concreto, informa decisiones sobre la fianza y la liberación condicional previa al juicio, la clasificación y ubicación en centros penitenciarios, y la determinación del

[14] Por instrumento actuarial (de evaluación del riesgo) se entiende una herramienta estructurada y estadística que, a partir del análisis de datos históricos de personas y casos previos, identifica un conjunto predeterminado de factores de riesgo y/o protectores, les asigna valores numéricos y los pondera/combina para producir un puntaje o categoría de riesgo (p. ej., bajo/medio/alto). Ese puntaje expresa una estimación probabilística basada en el comportamiento observado en el grupo de referencia con características similares, por lo que pretende informar la decisión (aportar un insumo empírico) y no "adivinar" con certeza el comportamiento de un individuo en particular.

[15] Por necesidades criminógenas se entienden aquellos factores dinámicos, tales como conductas, actitudes o condiciones, que están directamente vinculados a la conducta delictiva y, por tanto, a la probabilidad de reincidencia. En el modelo Riesgo–Necesidad–Respuesta (RNR), el principio de necesidad exige que la intervención se concentre en esas necesidades (las que aumentan el riesgo), y no en necesidades no criminógenas que pueden ser importantes para el bienestar, pero cuya relación con la reincidencia es débil o inexistente.

nivel de supervisión necesario en libertad vigilada o libertad condicional (Equivant, 2017; Dressel & Farid, 2018).

En síntesis, COMPAS se presenta como una herramienta de organización de información y estimación de riesgo para decisiones estatales que reparten cargas (vigilancia, restricción, encarcelamiento) y beneficios (alternativas a la prisión o libertad supervisada), sin asumirse formalmente como sustituto único del decisor humano (Equivant, 2017).

6.1.2. Contexto institucional, población y territorio impactado

COMPAS impacta principalmente a la población adulta vinculada al sistema de justicia penal en Estados Unidos. El software está diseñado para ser aplicado a hombres y mujeres que han sido recientemente retirados de la comunidad o que ya se encuentran bajo supervisión comunitaria, por lo que su alcance se proyecta sobre personas en distintas fases del proceso penal (Equivant, 2017).

En cuanto a su despliegue territorial, COMPAS no se circunscribe a una región única, sino que es utilizado por agencias de justicia penal en distintos estados de Estados Unidos. Su aplicación depende de la adopción institucional por parte de autoridades estatales o locales; por ello, el territorio impactado no responde a un espacio homogéneo, sino a los lugares donde el instrumento se incorpora a los flujos de decisión, con diferencias relevantes en protocolos, momento de uso y efectos prácticos (Angwin et al., 2016).

Ese impacto tampoco se distribuye de manera uniforme. La literatura advierte una mayor exposición a puntajes elevados para ciertos grupos: (i) minorías raciales, en la medida en que los datos históricos del sistema penal pueden trasladarse a clasificaciones de riesgo desiguales; (ii) hombres, cuando el diseño o el uso del instrumento incorpora diferencias por

sexo o género; y (iii) personas en contextos socioeconómicos desfavorecidos, dado que variables asociadas a marginalidad pueden ser captadas de forma directa o indirecta por el modelo (Mayson, 2018).

6.1.3. Datos: procedencia, construcción y riesgos de sesgo

El sistema COMPAS utiliza una estructura de datos híbrida, combinando (i) registros administrativos del sistema penal y (ii) entrevistas estructuradas y autorreporte del evaluado (Equivant, 2017). Esta arquitectura es metodológicamente relevante porque los registros oficiales pueden incorporar huellas históricas del policiamiento selectivo y de la criminalización diferencial, de modo que ciertas variables funcionen como proxies de categorías protegidas tales como raza, pobreza y territorio (Mayson, 2018; Washington, 2019).

En términos de pertinencia y representatividad, la documentación técnica reporta el uso de *norm groups* para transformar puntajes brutos en deciles comparativos, apoyados en datos normativos recolectados a partir de más de 30,000 evaluaciones realizadas entre 2004 y 2005 en prisiones, cárceles y agencias de libertad condicional de Estados Unidos (Equivant, 2017). Sin embargo, la representatividad y validez local de datos ha sido objeto de crítica académica y judicial recurrente; por ejemplo, la Corte Suprema de Wisconsin advirtió en el caso State v. *Loomis* (2016) que la herramienta no contaba, al momento de su empleo, con validación específica para la población local, lo que tensiona su desempeño en jurisdicciones concretas.

Sobre la calidad y cantidad de los datos existe una tensión documentada. Mientras la empresa desarrolladora defiende la necesidad de un conjunto amplio de factores para

mantener la precisión (Equivant, 2017), investigaciones independientes sostienen que un número menor de variables pueden aproximar resultados comparables, lo que abre un cuestionamiento metodológico sobre si el volumen de datos recolectado es estrictamente necesario (Dressel & Farid, 2018). Además, se ha señalado que COMPAS opera sobre la probabilidad de nuevo arresto más que sobre la comisión real de un delito, lo que puede introducir sesgos derivados de prácticas policiales diferenciales y afectar por tanto la pertinencia de las etiquetas de riesgo (Mayson, 2018; Washington, 2019).

Finalmente, aunque el diseño se presenta como formalmente "ciego a la raza", su dependencia de variables correlacionadas con la raza y el estatus socioeconómico, tales como el historial penal, la educación, el empleo o la estabilidad residencial, mantiene el riesgo de producir impactos diferenciados por vía de proxies (indicadores indirectos). Por ello, la evaluación del dato no puede agotarse en la ausencia de variables sensibles explícitas, sino que debe atender a la estructura correlacional y al origen social de los registros utilizados (Dieterich et al., 2016; Dressel & Farid, 2018; Mayson, 2018).

6.1.4. Tratamiento de datos y flujos de información

El tratamiento de datos en el ecosistema COMPAS inicia con una recolección que integra dos fuentes principales: datos administrativos extraídos de expedientes penales y registros públicos y, por otro lado, incorpora la auto información obtenida mediante un cuestionario respondido por el evaluado o completado durante una entrevista estructurada del cual se extraen 137 características que abarcan temas diversos tales como antecedentes familiares, educación, empleo y actitudes personales (Equivant, 2017). Un rasgo relevante del flujo es que la recolección de datos no se concibe como completamente rígida, pues se

prevé la posibilidad de corrección y complementación por parte del personal encargado cuando se detectan omisiones o errores en la información registrada (State v. Loomis, 2016).

De otro lado, COMPAS no funciona como una herramienta aislada, sino como un módulo dentro de la *Northpointe Suite*, un sistema web integrado de gestión de casos y evaluación. Ello permite el cruce de datos y la continuidad de la información entre agencias, de modo que los perfiles de riesgo y necesidades resultan accesibles en varias etapas del proceso penal, desde los servicios previos al juicio hasta la libertad condicional y el sistema penitenciario (Equivant, 2017). El sistema actúa así, como una extensión de los sistemas de información judicial existentes.

6.1.5. Modelo o arquitectura del sistema

COMPAS se ubica en la familia de herramientas actuariales: combina escalas y fórmulas para producir puntajes de riesgo y necesidades, con base en el conjunto de variables descrito en el apartado Datos. La documentación técnica lo presenta como un instrumento que calcula sus escalas mediante ecuaciones lineales (Equivant, 2017) y trabajos independientes han sostenido que su comportamiento puede aproximarse al de un clasificador lineal simple, alcanzando niveles de precisión comparables con un número menor de variables (Dressel & Farid, 2018).

En cuanto a las variables y características, el sistema procesa hasta 137 indicadores extraídos de registros oficiales y entrevistas estructuradas (Dressel & Farid). Estos se dividen en factores estáticos, como el historial criminal, y variables dinámicas, que responden a necesidades criminogénicas (State v. Loomis, 2016). Para la Escala de Riesgo de Reincidencia General, los insumos críticos incluyen el involucramiento criminal previo,

problemas vocacionales o educativos, historial de consumo de drogas, edad al momento de la evaluación y edad en el primer arresto. Por su parte, la Escala de Riesgo de Reincidencia Violenta integra el historial de violencia y de incumplimiento, junto con factores demográficos y socio-educativos. Es notable que el modelo excluye explícitamente el delito actual de su cálculo de probabilidad de reincidencia, enfocándose exclusivamente en el perfil histórico y contextual del individuo (Equivant, 2017).

6.1.6. Salida del sistema y alcance inferencial

El output o salida principal de COMPAS (Equivant, 2017) suele expresarse como clasificaciones o puntajes de riesgo en una escala de deciles (1–10) y agrupadas para uso operativo en tres categorías: bajo (1–4), medio (5–7) y alto (8–10).

El reporte suele representar visualmente el perfil de riesgo e incorpora, además, indicadores de necesidades criminógenas orientados a guiar decisiones de gestión de caso y planes de intervención (Equivant, 2017). En particular, el sistema reporta resultados para reincidencia general, reincidencia violenta y otros componentes de riesgo y gestión asociados al ciclo correccional, complementados con áreas de necesidad relevantes para tratamiento o supervisión (Dieterich et al., 2016).

El modelo presenta límites estructurales definidos por su propia naturaleza estadística. Al basarse en datos agregados, el sistema identifica grupos de riesgo, pero carece de la capacidad técnica para realizar predicciones a nivel individual con certeza. Debido a esto, se establece como un límite operativo que los puntajes no deben ser el factor determinante en la decisión judicial ni utilizarse para fijar la severidad de una sentencia, sino actuar como un soporte consultivo para el juicio profesional (State v. Loomis).

Estos puntajes son relativos al grupo de norma (norm group), por lo que expresan una ubicación comparativa respecto de una población de referencia y no una magnitud absoluta de riesgo, es decir, reflejan el riesgo del individuo en comparación con una población de referencia. Dicho en otras palabras, el alcance inferencial del resultado es estadístico y grupal: el puntaje deriva de patrones observados en grupos con características similares y no permite predicción individual con certeza (Equivant, 2017). Por ello, el debate judicial ha subrayado la necesidad de advertencias explícitas sobre el alcance del puntaje y sobre su lectura como insumo auxiliar, no como determinación individual concluyente (State v. Loomis, 2016).

6.1.7. Rol en la decisión estatal

El rol de COMPAS se define institucionalmente como un instrumento orientado a guiar clasificación, supervisión y planificación de tratamientos, sin sustituir el juicio del funcionario (Equivant, 2017). En la práctica, su salida se integra habitualmente como anexo del Informe de Investigación Presentencia (PSI), funcionando como insumo consultivo para el juez en la fase de sentencia (Washington, 2019; State v. Loomis, 2016,). En esa etapa, el puntaje se usa para valorar si una persona puede ser supervisada de manera segura en comunidad o si requiere encarcelamiento, lo cual explica su peso real dentro del circuito decisional (Harvard Law Review, 2017; State v. Loomis, 2016).

En el plano formal, COMPAS se presenta como un insumo de apoyo. Sin embargo, el riesgo metodológico reside en su desplazamiento práctico: aunque se afirme que “no es determinante”, el puntaje puede operar como refuerzo de decisión o como una prueba técnica difícil de controvertir cuando ingresa a decisiones restrictivas. Precisamente por ello, State

v. Loomis (2016) admite su uso solo bajo cautelas: que no sea el único fundamento, que se adviertan su naturaleza y límites inferenciales, y que se reconozca que el instrumento no debe decidir por sí mismo la sentencia.

Aun con esas salvaguardas, persiste un efecto de anclaje, y es que el decisor, debido a sesgos cognitivos, puede otorgar un peso desproporcionado al output algorítmico por su apariencia experta, especialmente en contextos mediados por una fuerte presión laboral en búsqueda de una mayor eficiencia, donde además existe una alta carga administrativa. En ese sentido, aunque normativamente se afirme que el algoritmo no sustituye el juicio profesional, la promesa de eficiencia puede incentivar una dependencia práctica de sistemas como COMPAS para gestionar grandes volúmenes de casos (Washington, 2019, p. 159).

6.1.8. Gobernanza institucional y marco jurídico declarado

La gobernanza del sistema COMPAS se estructura bajo un modelo de colaboración público-privada, donde la empresa Equivant funge como desarrolladora y propietaria del activo digital, mientras que diversas agencias estatales operan la herramienta dentro de sus rutinas administrativas y de decisión. En la cadena de mando institucional, el personal de correcciones es el responsable directo de generar los reportes que integran los puntajes de riesgo, los cuales son posteriormente entregados a los jueces para guiar la toma de decisiones (State v. Loomis,) lo que sitúa el sistema dentro del flujo real de adjudicación y gestión del castigo. Esta transición hacia herramientas comerciales responde a una necesidad de optimizar la eficiencia administrativa frente al crecimiento masivo de la población carcelaria, desplazando en muchos casos a los sistemas de evaluación desarrollados internamente por el Estado (Washington, 2019).

En materia de evaluación y control, el sistema ha sido objeto de diversos estudios. Algunos estados, como Nueva York, han publicado estudios de validación que reportan niveles satisfactorios de precisión, mientras investigaciones independientes, en particular la de ProPublica, han operado como auditorías externas al denunciar disparidades relevantes en las tasas de error según la raza (Angwin et al., 2016a). A ello se suma el control jurisdiccional: la Corte Suprema de Wisconsin ordenó que cada PSI incluya una advertencia escrita sobre las limitaciones del software y la ausencia de validación local específica (State v. Loomis, 2016).

No obstante, persiste una crítica central, el secreto comercial impide una auditoría plena de la lógica que subyace a la clasificación de las personas y dificulta delimitar con claridad las responsabilidades por sus efectos (Washington, 2019).

6.1.9. Transparencia, trazabilidad y auditabilidad

La transparencia del sistema COMPAS se encuentra severamente restringida por su naturaleza privada, lo cual constituye una barrera fundamental para la audibilidad plena en el marco estatal. La empresa desarrolladora invoca la protección de secretos comerciales para no divulgar plenamente los procedimientos computacionales internos, el código fuente, ni la ponderación específica de las variables que conforman el algoritmo (State v. Loomis, 2016), lo cual limita el escrutinio público y el control contradictorio cuando el puntaje ingresa a decisiones restrictivas.

Esta configuración técnica genera un régimen de opacidad donde, si bien es posible realizar auditorías externas de impacto sobre los resultados, la lógica subyacente que

determina la clasificación de un individuo permanece como una *caja negra* para todos los actores del proceso judicial (Angwin et al., 2016a).

En materia de trazabilidad, el sistema permite cierto control sobre los insumos, pero no del proceso de transformación. Los acusados tienen acceso al reporte final y al cuestionario de 21 preguntas clave, lo que les permite verificar y corregir la exactitud de los datos de entrada (State v. Loomis, 2016). Sin embargo, la explicabilidad para el afectado es baja, y es que aunque puede refutar sus respuestas, carece de mecanismos para comprender cómo se derivó su puntuación o por qué fue comparado con un grupo normativo específico (Washington, 2019).

6.1.10. Control humano significativo

El CHS existe en varios puntos del ciclo de vida de COMPAS, desde la captura de datos, interpretación y decisión final, pero su efectividad depende de dos condiciones: la capacidad real de apartarse del output y la capacidad real de comprenderlo y controvertirlo. State v. Loomis (2016) resulta ilustrativo porque fija un criterio mínimo: el puntaje no debe ser determinante y el decisor debe tratarlo como apoyo con límites, no como autoridad concluyente. Con todo, si el sistema es opaco y el operador carece de herramientas para comprender tasas de error y sesgos por subgrupos, el control humano en consecuencia, corre el riesgo de volverse meramente formal.

Este control se manifiesta en tres etapas críticas del ciclo de vida del sistema: primero en la supervisión ex ante y diseño, el control reside en la configuración institucional de la herramienta. Los usuarios y agencias tienen la facultad de configurar el sistema según los puntos de decisión específicos, seleccionando qué escalas aplicar (por ejemplo, escalas de

libertad condicional frente a escalas de reincidencia general) dependiendo de las necesidades administrativas y la población evaluada (Equivant, 2017). Asimismo, la validez del sistema depende de la supervisión humana experta para la creación y mantenimiento de los grupos de norma, asegurando que los datos de referencia sean representativos de la población local.

Durante la operación del sistema, la intervención humana es indispensable para la recolección y validación de los insumos. El personal de correcciones debe realizar entrevistas estructuradas y revisar expedientes penales para alimentar el algoritmo (State v. Loomis). Existe como se mencionó con anterioridad, un mecanismo de control de calidad donde el evaluado tiene la posibilidad de refutar, explicar o suplementar la información si se detectan inexactitudes en las respuestas que alimentan el cálculo del riesgo (State v. Loomis, 2016). La interpretación de la salida también requiere que el profesional conecte todos estos puntos con el puntaje actuarial, las teorías criminológicas y la historia única del individuo (Equivant, 2017).

Finalmente, en la revisión ex post y la posibilidad de apartarse, el sistema integra formalmente el juicio profesional a través del "*override*" o anulación. La documentación oficial establece que se espera que el personal humano esté en desacuerdo con la puntuación algorítmica en aproximadamente el 10% de los casos debido a circunstancias mitigantes o agravantes que el software no puede capturar (Equivant, 2017). En el ámbito jurisdiccional, la sentencia Loomis (2016) reforzó este control al determinar que el juez no puede utilizar el puntaje como el factor determinativo para la encarcelación o la severidad de la pena, y tiene la obligación de articular factores independientes que sustenten su decisión. No obstante, la doctrina advierte sobre el riesgo del sesgo de automatización, donde la celeridad

administrativa y la confianza en los números pueden reducir este control humano a una función meramente decorativa (Washington, 2019).

6.1.11. Evidencia de impactos: afectaciones y mecanismos de contención

La evidencia de impactos de COMPAS se volvió paradigmática por el debate sobre disparidades en las tasas de error y por la discusión metodológica acerca de qué debe entenderse por sesgo cuando se emplean distintas métricas de equidad. Una formulación sintética del problema aparece en la réplica publicada en Federal Probation, que reproduce la tesis central atribuida a ProPublica: el algoritmo se equivoca de manera distinta entre acusados negros y blancos, con una tendencia a clasificar erróneamente a personas negras como de alto riesgo con mayor frecuencia relativa (Flores et al., 2016).

El hallazgo más citado indica que los acusados negros que no reincidieron fueron clasificados erróneamente como de "alto riesgo" (falsos positivos) en una tasa del 44.9%, frente al 23.5% de los acusados blancos. A la inversa, los acusados blancos que sí reincidieron fueron clasificados erróneamente como de "bajo riesgo" (falsos negativos) con mayor frecuencia que los negros: 47.7% frente a 28.0% (Angwin et al., 2016a).

Desde la perspectiva del desarrollador y de sectores académicos afines, esas diferencias no demostrarían un sesgo intrínseco, sino un resultado matemático asociado a distintas tasas base de reincidencia en la población analizada (Flores et al., 2016). Bajo esa lógica, el sistema cumpliría con la paridad predictiva, pues una puntuación de "alto riesgo" representaría una probabilidad similar de reincidencia con independencia de la raza. Sin embargo, se ha objetado que el uso del arresto como variable proxy del delito proyecta sobre el algoritmo las desigualdades históricas de la vigilancia policial (Mayson, 2018).

En conjunto, la evidencia disponible muestra una tensión persistente entre la precisión técnica declarada y los impactos diferenciados sobre minorías raciales, hombres y personas en contextos de desventaja socioeconómica, especialmente cuando variables asociadas a pobreza, empleo o estabilidad residencial inciden en la clasificación del riesgo (Washington, 2018).

Estos impactos también se han traducido en litigios y ajustes institucionales. El caso State v. Loomis (2016) consolidó la preocupación por la opacidad derivada del secreto comercial y estableció que, aunque el uso de COMPAS no vulnera per se el debido proceso, los jueces deben recibir advertencias obligatorias sobre sus limitaciones, la falta de validación local y el riesgo de disparidades raciales. A ello se suma el efecto de anclaje, que puede desplazar la valoración individualizada del caso en favor del puntaje numérico (Harvard Law Review, 2017).

6.1.12. Síntesis del caso

Tabla 4
Síntesis de la caracterización técnico-institucional del caso COMPAS

Dimensión	Preguntas guía	Hallazgos clave
Finalidad declarada y contexto	¿Qué problema pretende resolver? ¿En qué decisión estatal se inserta?	COMPAS es un sistema actuarial de evaluación de riesgo y necesidades diseñado para apoyar decisiones del sistema penal y correccional en EE. UU. Su finalidad declarada es organizar información para estimar riesgo de reincidencia y necesidades criminógenas, con miras a mejorar consistencia y eficiencia en decisiones como supervisión, clasificación, manejo de casos y, en ciertos contextos, informes pre-sentenciales.

Dimensión	Preguntas guía	Hallazgos clave
Población y territorio impactado	¿A quién aplica? ¿Se focaliza territorial o poblacionalmente?	Impacta principalmente a personas adultas vinculadas al sistema penal en sus diferentes etapas. Su despliegue es territorialmente variable, dependiendo de adopción por agencias estatales y locales; por eso su efecto real depende de la jurisdicción, la fase procesal y el uso institucional concreto. Los grupos más expuestos suelen ser poblaciones sometidas a mayor vigilancia penal (comunidades racializadas y territorios empobrecidos), por el peso de historiales y factores socio-contextuales.
Datos: origen y calidad	¿De dónde vienen los datos? ¿Son pertinentes, representativos, actualizados?	Los datos provienen de una estructura híbrida: registros oficiales (historial penal, arrestos, condenas, edad, etc.) y entrevista o autorreporte estandarizado para necesidades criminógenas. Aunque el diseño omite la raza como variable explícita, incluye múltiples proxies (historial policial-penal, empleo, educación, estabilidad residencial) que pueden trasladar desigualdades estructurales al puntaje. La calidad y representatividad de estos datos depende del grupo normativo y de validaciones aplicables al contexto local; la selección del objetivo (p. ej., nuevo arresto) puede amplificar sesgos ligados a vigilancia diferencial.
Tratamiento de datos	¿Cómo se recolectan, cruzan, conservan y comparten?	El tratamiento integra recolección por personal correccional (entrevista y registros), consolidación en plataforma de gestión de casos y generación de reportes para uso institucional. Hay posibilidad de corregir o refutar algunos insumos, pero la información pública suele ser insuficiente para reconstruir con rigor retención, interoperabilidad detallada, acceso interno y sus respectivos protocolos.
Modelo/arquitectura	¿Qué tipo de modelo (reglas/ML/DL)? ¿Qué variables usa?	COMPAS opera como herramienta actuarial basada en escalas: usa múltiples variables (hasta 137 reportadas) para

Dimensión	Preguntas guía	Hallazgos clave
		producir puntajes de riesgo y perfiles de necesidades. Aunque se conocen rasgos generales (factores estáticos o dinámicos y normalización por grupos), una parte relevante del funcionamiento es reservada (caja negra): no hay acceso público completo al detalle de ponderaciones, entrenamiento y calibración por jurisdicción, lo que es jurídicamente sensible cuando el puntaje entra en decisiones restrictivas.
Salida del sistema	¿Qué produce (score, alerta, recomendación)?	La salida principal son puntajes y categorías de riesgo (usualmente expresados en deciles y agrupables en bajo/medio/alto) para reincidencia general y violenta, más indicadores de necesidades criminógenas. El alcance inferencial es estadístico y grupal: el puntaje describe tendencias relativas frente a una población de referencia y no equivale a certeza individual; por ello existen advertencias institucionales y judiciales sobre su interpretación y límites.
Rol en la decisión	¿La salida decide, recomienda o prioriza?	Formalmente COMPAS se presenta como insumo de apoyo para orientar clasificación, supervisión y manejo de casos. En la práctica, puede producir efecto de determinación indirecta: el puntaje funciona como anclaje o prueba técnica difícil de controvertir, reforzando decisiones ya inclinadas por la automatización, especialmente en escenarios de alta carga administrativa.
Gobernanza, evaluación de impacto y responsabilidades	¿Quién opera? ¿Quién responde? ¿Qué controles y evaluaciones existen?	La gobernanza es fragmentada: Equivant desarrolla y actualiza; agencias usuarias administran el instrumento e integran reportes; el decisor estatal (juez o autoridad) conserva responsabilidad formal. Los controles aparecen de forma desigual (validaciones técnicas, monitoreo o litigios relevantes), más como gobernanza reactiva/ex post que

Dimensión	Preguntas guía	Hallazgos clave
		como esquema integral ex ante comparable a una evaluación de impacto en derechos uniformemente exigida.
Transparencia y trazabilidad	¿Es auditable? ¿Hay explicabilidad/registro?	La transparencia es limitada por secretos comerciales: no se divulgan plenamente código o lógica interna, lo que reduce la auditabilidad sustantiva. Existe trazabilidad documental del reporte final (p. ej., en PSI) y cierto control de insumos (revisión de respuestas o datos básicos), pero la explicación al afectado sobre por qué se asignó un riesgo concreto es baja; la trazabilidad del proceso interno es parcial.
Control humano significativo (CHS)	¿Dónde interviene el humano (diseño-uso-revisión)?	El humano interviene en (i) adopción y diseño institucional (definición de usos), (ii) operación (entrevista, carga y corrección de datos, lectura del reporte) y (iii) revisión y decisión (posibilidad formal de apartarse). Sin embargo, el CHS se debilita cuando la opacidad impide comprender y controvertir el fundamento del puntaje, y cuando el puntaje opera como anclaje en la práctica.
Evidencia de impactos	¿Hay evidencia de sesgos, errores o afectaciones?	Existe evidencia documentada de impactos controvertidos: se han reportado disparidades en tipos de error (falsos positivos y negativos por raza y sexo) y abrió discusión sobre qué métrica de equidad es relevante. Además, se cuestiona la proporcionalidad del tratamiento masivo de variables cuando modelos simples pueden aproximar rendimientos predictivos, y el caso judicial Loomis visibiliza tensiones de debido proceso ligadas a opacidad y uso en la sentencia.

6.2 Caso SyRI (Países Bajos)

6.2.1. Finalidad declarada y decisión estatal afectada

SyRI (Systeem Risico Indicatie) fue concebido como un instrumento legal y técnico para prevenir y combatir el uso indebido de recursos públicos y diversas formas de fraude vinculadas al sistema de protección social en Países Bajos. En términos normativos, se desarrolló con el objeto de ser usado en el procesamiento y análisis de datos a solicitud de entidades públicas competentes con fines de análisis de riesgo en materias como prestaciones, subsidios, impuestos y otros esquemas conexos. En el discurso institucional, esa finalidad se presenta como una respuesta "racionalizadora": permitir una fiscalización más eficiente del fraude en el sostenimiento del sistema social (Rechtbank Den Haag, 2020)

Metodológicamente, la decisión estatal afectada por SyRI no es (al menos en su formulación formal) una sanción automática, sino la decisión de selección y priorización de casos: SyRI buscaba identificar personas (o entidades) con "mayor probabilidad" de fraude (Rechtbank Den Haag, 2020) y producir un reporte de riesgo que habilitaba o intensificaba actuaciones posteriores de verificación e investigación por autoridades centrales y locales (Meuwese, 2020).

Por eso, el punto clave no es "qué castigo impone", sino cómo redistribuye el umbral de sospecha y a quién coloca en el circuito institucional de control (inspección, requerimientos, cruces adicionales, apertura de indagaciones), con potencial incidencia ulterior en decisiones sobre continuidad de beneficios o recuperación de recursos, pero siempre como derivación posterior (Alston, 2019). En cuanto al contexto de uso, SyRI se insertó en prácticas de cooperación interinstitucional y analítica de datos en donde

autoridades públicas compartían y analizaban información antes fragmentada en diversos espacios.

6.2.2. Contexto institucional, población y territorio impactados

SyRI se dirigió, en términos poblacionales, a personas asociadas a la asistencia social y, en general, a residentes cuyo riesgo de "conducta fraudulenta" se pretendía detectar mediante el cruce masivo de datos por parte del Estado y gobiernos locales de Países Bajos (Wieringa, 2023). Su impacto no se configuró como "universal" sobre toda la población, sino como selectivo: el propio diseño institucional del despliegue lo ubicó en contextos de beneficiarios de bienestar social y control antifraude, en los cuales la salida del sistema (reportes de riesgo) podía traducirse en investigaciones por parte de autoridades administrativas.

Territorialmente la evidencia disponible indica una focalización por zonas: SyRI fue utilizado exclusivamente en vecindarios con tasas significativas de pobreza, criminalidad, desempleo y beneficiarios de asistencia, a los que el propio Estado llegó a denominar "barrios problemas" (Wieringa, 2023). El despliegue operaba mediante proyectos locales y municipales.

Bajo esa forma de implementación, los grupos más expuestos fueron, precisamente, quienes habitan esos territorios priorizados y quienes se encontraban más conectados a circuitos de asistencia social, personas de bajos ingresos, así como migrantes y minorías étnicas, por lo cual la herramienta fue criticada por emplearse exclusivamente en áreas con alta proporción de estos grupos, con impacto particularmente gravoso sobre los más pobres (Appelman et al., 2021).

En consecuencia, el rasgo relevante no es solo "a quién aplica", sino que la selección territorial de los barrios "problema" funcionó como criterio estructurante de exposición, intensificando riesgos de estigmatización territorial y de discriminación indirecta por concentración de vigilancia administrativa (Wieringa, 2023).

6.2.3. Datos: procedencia, construcción y riesgos de sesgo

En cuanto al origen de los datos, SyRI se alimenta de un acoplamiento interinstitucional de registros administrativos ya existentes en distintas entidades públicas, que son enlazados y agregados para su procesamiento algorítmico. La base normativa del sistema enumeró de manera exhaustiva las categorías de datos que podían ser objeto de tratamiento en proyectos SyRI: trabajo e ingresos; datos laborales; datos tributarios; datos sobre vivienda; educación; pensiones, beneficios y prestaciones; seguros de salud; información sobre demandas y reclamaciones; multas y medidas administrativas; deudas; datos de identificación; datos en relación con la economía; datos derivados del cumplimiento de la legislación laboral; registros de asistencia escolar; la base municipal de datos personales; y el registro de vehículos (Raad van State, 2014).

Sobre la calidad y representatividad de los datos, el punto metodológicamente decisivo es que estos reutilizan registros administrativos creados para fines diversos (como tributación o vivienda), lo que puede distorsionar la realidad social que el algoritmo intenta analizar. Durante su proceso de regulación, se advirtió que la formulación de categorías era tan amplia que permitía recopilar información no solo de sospechosos, sino también de familiares y otros convivientes, lo que vulnera el criterio de necesidad y multiplica la posibilidad de errores por el arrastre de datos asociados (Raad van State, 2014).

Por otro lado, aunque el régimen declaraba operar con categorías cerradas, en la práctica el volumen de datos disponibles era tan vasto que las salvaguardas de minimización y limitación de finalidad resultaban ineficaces ante la opacidad del modelo (Rachovitsa & Johann, 2022). Al utilizar indicadores sobre residencia, deudas o salud, el sistema emplea proxies o variables sustitutas que reflejan el nivel socioeconómico o el origen migratorio de las personas; esto facilita una discriminación indirecta al inferir riesgos a partir de patrones de control históricamente sesgados (Rachovitsa & Johann, 2022)

6.2.4. Modelo o arquitectura del sistema

Desde el punto de vista técnico, SyRI operaba como un sistema de perfilamiento de riesgo: una vez enlazados distintos conjuntos de datos administrativos de entidades públicas, el agregado era procesado por un modelo de riesgo que buscaba detectar patrones asociados a presunto fraude en prestaciones sociales y ámbitos conexos. El resultado no se presentaba como una explicación del caso individual, sino como una señalización que podía culminar en la emisión de un reporte de riesgo para que las autoridades competentes iniciaran verificaciones o investigaciones posteriores (Rachovitsa & Johann, 2022; Wieringa, 2023).

6.2.5. Salida del sistema y alcance inferencial

La salida principal de SyRI no era una sanción automática ni una decisión administrativa definitiva, sino una señalización de riesgo. En términos operativos, el tratamiento culminaba —cuando el modelo identificaba un caso como relevante— en la generación de una reporte o alerta de riesgo dirigida a las autoridades competentes, que servía como argumento de priorización para actuaciones posteriores de verificación o investigación. Esa alerta se incorporaba, además, a un registro de reportes de riesgo, desde el cual podía ser

consultada y comunicada a los participantes en la medida necesaria para el cumplimiento de sus funciones legales (Staatsblad, 2014, p. 4).

En ese sentido, el output puede caracterizarse como una alerta con valor administrativo: SyRI producía un resultado que clasifica o prioriza (alto riesgo vs. no alto riesgo), habilitando el paso de la persona o entidad desde un universo amplio de datos cruzados hacia un circuito más intenso de control institucional (Staatsblad, 2014).

El punto central sobre alcance inferencial es que esa alerta no se presenta como prueba del fraude, sino como indicio algorítmico que orienta recursos de investigación. Sin embargo, el carácter de insumo no elimina su carga decisional: al reconfigurar el umbral de sospecha y activar investigaciones en poblaciones específicas, el output tiene capacidad de producir efectos materiales relevantes (por ejemplo, apertura de verificación, requerimientos, controles intensificados), aun cuando formalmente se ubique antes de una decisión final sobre beneficios o sanciones (Rachovitsa & Johann, 2022). En ese marco, la dificultad jurídica radica en que el sujeto afectado normalmente no dispone de una explicación verificable sobre por qué fue marcado como riesgo, ni de la lógica interna que conduce a esa clasificación, lo que impacta la posibilidad de control contradictorio del output (Rachovitsa & Johann, 2022).

6.2.6. Rol en la decisión estatal

En el plano institucional y práctico, el rol de SyRI se vuelve más intenso: el resultado del puntaje de riesgo opera como gatillo para la investigación posterior por parte de las administraciones públicas que integran el proyecto, es decir, funciona como un filtro de priorización que orienta qué casos merecen ser investigados con recursos de control estatal

(Bekker, 2021). En esa medida, aunque formalmente sea un insumo, su output tiende a tener efectos de determinación indirecta: decide qué entra al circuito de investigación reforzada y qué queda fuera, con el riesgo de que la notificación sea tratada como una razón suficiente para intensificar vigilancia o control, especialmente cuando el sujeto no recibe información individualizada y no puede conocer oportunamente por qué fue seleccionado (Staatsblad, 2014; Bekker, 2021).

6.2.7. Gobernanza institucional y marco jurídico declarado

La gobernanza formal de SyRI se diseña como un esquema centralizado de responsabilidad ministerial articulado con proyectos solicitados por sistemas de cooperación interinstitucional. En términos normativos, la lógica es: una coalición de entidades solicita el uso de SyRI, y si el requerimiento cumple condiciones (propósito concreto, organización del proyecto, duración, participantes y datos a tratar), el Ministro de Asuntos Sociales y Empleo asume la responsabilidad de procesar los datos en SyRI y de conducir el proyecto dentro de un marco reglamentario específico (Staatsblad, 2014).

En la cadena operativa, la arquitectura institucional distribuye funciones entre (i) un procesador técnico y (ii) una unidad de análisis estatal.

6.2.8. Transparencia, trazabilidad y auditabilidad

En términos de auditabilidad, el principal problema de SyRI es que el marco normativo no ofrece un acceso verificable a la lógica del modelo de decisión, es decir, no se especifica públicamente cómo funciona el modelo, ni cuáles son (o pueden ser) los indicadores utilizados en cada proyecto. Esto limita de manera estructural la posibilidad de

control externo, porque ni la ciudadanía ni los afectados pueden reconstruir con precisión por qué un conjunto de datos termina en una señal de riesgo (Rechtbank Den Haag, 2020). En la misma línea, el tribunal de la Haya subrayó que la normativa tampoco brinda información suficiente sobre la validación del modelo y de los indicadores; incluso en el proceso judicial no se contaba con el nivel de detalle necesario para controlar cómo se llega a los resultados, lo cual dificulta que una persona se defienda frente al informe de riesgo (Rechtbank Den Haag, 2020).

Finalmente, en explicabilidad para el afectado, el estándar aparece debilitado: no hay un esquema robusto de explicación del resultado (reglas y ponderaciones) y el propio tribunal resaltó que el derecho al respeto por la vida privada implica que el afectado pueda seguir sus datos en medida razonable, exigencia que se ve comprometida cuando no hay transparencia suficiente sobre la operación del modelo y sus criterios (Rechtbank Den Haag, 2020).

6.2.9. Control humano significativo

En la operación, SyRI incorpora intervención humana pero con una arquitectura de dos fases: (i) una fase 1 predominantemente técnica y automatizada a cargo del operador para producir una preselección defraudes potenciales; y (ii) una fase 2 donde esos potenciales hallazgos se descifran y son analizados por la unidad de analis, que depura coincidencias irrelevantes y, con base en la selección final, emite las informes de riesgo (Staatsblad, 2014). Este punto es clave puesto que la decisión que activa la consecuencia institucional está mediada por análisis humano, pero sobre un universo previamente filtrado por el modelo automático (Staatsblad, 2014).

En la revisión ex post, el diseño normativo prevé dos controles relevantes: primero, la retroalimentación obligatoria por parte de los receptores de los informes de riesgo (si se siguieron o no, resultados y utilidad), dentro de un plazo (hasta veinte meses), con la finalidad explícita de mejorar la efectividad y eventualmente ajustar el modelo de riesgo; segundo, reglas de destrucción y retención, según las cuales el operador destruye en plazos definidos sus archivos del proyecto y los datos de personas que no se convierten en informes de riesgo (Staatsblad, 2014).

Ahora bien, para determinar si el control humano es significativo y no solo una firma humana sobre resultados opacos, la reconstrucción debe registrar una limitación estructural, la intervención humana ocurre, pero con baja transparencia externa del modelo y con déficits de acceso e impugnación para el sujeto. El propio diseño del registro permite que una persona solicite información sobre si está incluida, pero la solicitud puede ser rechazada si hay investigación en curso, justamente para no revelar el modus operandi (Staatsblad, 2014). Sumado a ello, análisis doctrinales destacan que la opacidad intencional en sistemas algorítmicos públicos dificulta el ejercicio efectivo de derechos y debilita el control judicial, llegando incluso a señalar que, en SyRI, la falta de transparencia inhibía a las personas para ejercer sus derechos y afectaba la capacidad de los tribunales de ejercer escrutinio adecuado (Rachovitsa & Johann, 2022).

6.2.10. Evidencia de impactos: afectaciones y mecanismos de contención

La evidencia disponible sobre SyRI permite reconstruir (i) afectaciones documentadas y riesgos materializados, y (ii) mecanismos de contención activados (principalmente por vía judicial). A diferencia de COMPAS, en SyRI no existe evidencia

pública robusta y verificable sobre tasas de error en materia de falsos positivos del modelo de riesgo, precisamente porque el modelo y su lógica operativa permanecieron insuficientemente transparentes, lo cual representa un dato de impacto institucional: la imposibilidad de auditoría ciudadana plena y de control contradictorio sobre resultados que activan investigaciones.

En cuanto a afectación y disparidades, la evidencia más sólida proviene del fallo que declaró ilícito el marco de SyRI por vulneración del art. 8 CEDH, al considerar que el instrumento, en su forma vigente, no ofrecía salvaguardas suficientes ni permitía un balance justo entre el interés público (lucha contra fraude) y la injerencia en la vida privada, subrayando la responsabilidad especial del Estado al aplicar nuevas tecnologías (Meuwese, 2020; Rachovitsa & Johann, 2022). A ello se suma evidencia contextual ampliamente citada: SyRI fue desplegado en "barrios problema" y dirigido a zonas con mayores tasas de pobreza y beneficiarios, lo que incrementó el riesgo de discriminación indirecta por estatus socioeconómico o migratorio (Wieringa, 2023; ACNUDH, 2020).

En términos de contención, el impacto se concreta en una respuesta jurisdiccional fuerte: la decisión judicial detuvo el uso de SyRI en su forma vigente y, según análisis doctrinal, el Gobierno no apeló, con lo cual el sistema dejó de usarse por su forma insuficientemente transparente" (Meuwese, 2020, pp).

6.2.11. Síntesis del caso

Realizada la reconstrucción técnico-institucional del caso, a continuación, se presenta una síntesis de sus hallazgos conforme a la metodología de la Tabla 1. Su función es condensar, en un formato homogéneo, los elementos relevantes del sistema analizado —

finalidad, población impactada, datos, tratamiento, modelo, salida, rol, gobernanza, transparencia y trazabilidad, control humano significativo y evidencia de impactos— para visualizar de manera integrada cómo se articula el dispositivo y dónde se ubican sus principales puntos de tensión.

Tabla 5
Síntesis de la caracterización técnico-institucional del caso SyRI

Dimensión	Preguntas guía	Hallazgos clave
Finalidad declarada y contexto	¿Qué problema pretende resolver? ¿En qué decisión/incidencia estatal se inserta?	SyRI fue un sistema estatal de detección de fraude en el ámbito de seguridad social, concebido para apoyar la focalización de controles mediante perfilamiento de riesgo. Su finalidad declarada es mejorar la eficiencia al identificar señales de riesgo a partir de cruces masivos de datos administrativos, en proyectos coordinados entre distintas autoridades.
Población y territorio impactado	¿A quién aplica? ¿Se focaliza territorial o poblacionalmente?	Impacta a residentes y potenciales beneficiarios o usuarios de programas sociales, especialmente en zonas seleccionadas para proyectos de "riesgo" (focalización territorial). La selección por áreas y perfiles tiende a concentrar escrutinio en barrios vulnerables, lo que incrementa riesgos de estigmatización y vigilancia intensificada sobre grupos socioeconómicamente desfavorecidos.
Datos: origen y calidad	¿De dónde vienen los datos? ¿Son pertinentes, representativos, actualizados?	SyRI se alimentaba de múltiples bases administrativas inter-agenciales (p. ej., datos sobre beneficios, trabajo o ingresos, vivienda, impuestos y otros registros estatales), cuya cobertura suele ser amplia pero no necesariamente

Dimensión	Preguntas guía	Hallazgos clave
		neutral: los datos reflejan prácticas previas de control y pueden introducir proxies de pobreza, origen migratorio o territorialidad. La representatividad y calidad depende de la completitud de registros y de la validez de la etiqueta de fraude (riesgo de sesgo por selección: se observa más donde se controla más).
Tratamiento de datos	¿Cómo se recolectan, cruzan, conservan y comparten?	El tratamiento consistía en el cruce y análisis de datasets provenientes de distintas entidades para producir una señal de riesgo. El flujo incluía integración, procesamiento y circulación de reportes entre los participantes del proyecto. Aunque el marco regulaba aspectos de retención, destrucción y comunicación, el detalle operativo del cruce y de los accesos no era plenamente transparente, lo que convierte esa opacidad en un hallazgo relevante de gobernanza.
Modelo/arquitectura	¿Qué tipo de modelo (reglas/ML/DL)? ¿Qué variables usa?	La arquitectura funciona como un modelo de perfilamiento de riesgo (reglas y ML no claramente auditables públicamente) cuyo núcleo operativo se mantuvo bajo un régimen de secreto y limitación informativa. El rasgo estructural clave es la "caja negra": se conocen insumos generales y el objetivo (señal de riesgo), pero no es accesible con precisión el peso de variables ni la lógica interna, lo que compromete control democrático y contradicción.
Salida del sistema	¿Qué produce (score, alerta, recomendación)?	El output es una señal o reporte de riesgo (alerta) que prioriza casos para revisión posterior; no impone automáticamente una sanción, pero orienta investigaciones y controles. Su alcance inferencial es probabilístico e indicativo, con riesgo

Dimensión	Preguntas guía	Hallazgos clave
		de sobrerrepresentar alertas en contextos donde el sistema concentra la vigilancia.
Rol en la decisión	¿La salida decide, recomienda o prioriza?	SyRI operó como herramienta de priorización de fiscalización: define a quién se investiga primero y dónde se concentran recursos de control. Aunque formalmente es un "insumo", en la práctica puede producir automatización de facto al dirigir el escrutinio institucional y condicionar decisiones subsiguientes (aperturas de investigación, verificaciones y posibles efectos sobre acceso y continuidad de beneficios).
Gobernanza, evaluación de impacto y responsabilidades	¿Quién opera? ¿Quién responde? ¿Qué controles y evaluaciones existen?	La gobernanza es inter-agencial: autoridades nacionales y municipales coordinan proyectos y comparten insumos; la responsabilidad se fragmenta entre quien diseña el sistema, quien opera el cruce y quien actúa sobre el resultado. La experiencia SyRI muestra déficits en salvaguardas ex ante (evaluación de impacto y controles adecuados) y un control correctivo ex post vía tribunal, que terminó por cuestionar la legitimidad del esquema.
Transparencia y trazabilidad	¿Es auditable? ¿Hay explicabilidad/registro?	La transparencia fue considerada insuficiente: baja explicabilidad al afectado y limitaciones para auditar la lógica del perfilamiento. La trazabilidad del por qué un individuo o área resultaba marcada como riesgosa no era robusta ni accesible públicamente, lo que debilitó el control ciudadano y judicial sobre el sistema.
Control humano significativo (CHS)	¿Dónde interviene el humano (diseño-uso-revisión)?	El humano interviene en el diseño de proyectos, operación institucional y decisión de seguimiento

Dimensión	Preguntas guía	Hallazgos clave
		investigativo; sin embargo, el CHS es insuficiente si la intervención humana no puede comprender y explicar el criterio que disparó la alerta y si el afectado no cuenta con información para controvertirla. En SyRI, la falta de transparencia redujo la efectividad del control humano como garantía.
Evidencia de impactos	¿Hay evidencia de sesgos, errores o afectaciones?	El impacto documentado central es jurisdiccional: el tribunal de la Haya (2020) concluyó que el régimen SyRI vulneraba garantías (especialmente privacidad y vida privada) por falta de salvaguardas y proporcionalidad, con riesgo de estigmatización y discriminación indirecta al focalizar territorios vulnerables. El caso se convirtió en precedente comparado sobre límites al perfilamiento estatal en el llamado Estado de Bienestar.

6.3 Caso AFR Locate (Reino Unido)

6.3.1. Finalidad declarada y decisión estatal afectada

AFR Locate se presentó como una herramienta de reconocimiento facial en vivo orientada a mejorar la eficacia operativa de la policía en entornos públicos de alta afluencia (por ejemplo, eventos o zonas concurridas), mediante la localización rápida de personas previamente incluidas en una watchlist.

En ese marco, su problema objetivo se formula en términos de seguridad pública y optimización de recursos: permitir que la autoridad identifique con mayor rapidez a sujetos de interés (personas buscadas) y priorice su actuación en terreno sin depender exclusivamente de reconocimientos manuales o de búsquedas posteriores. Esta finalidad

aparece tanto en la descripción institucional del despliegue (South Wales Police, 2025) como en la evaluación independiente que caracteriza el sistema como un componente de apoyo a la identificación policial en espacios abiertos, diseñado para operar de manera abierta y con un objetivo de localización en tiempo real (Davies et al., 2018).

En cuanto a la decisión e incidencia estatal, AFR Locate se inserta en el momento operativo previo o inmediato a la intervención policial: el sistema no decide por sí mismo una medida jurídica, pero incide en decisiones estatales concretas como a quién detener o abordar, a quién requerir identificación, y eventualmente a quién conducir a verificación o arresto, en la medida en que su salida (alerta de posible coincidencia) actúa como disparador para una actuación policial posterior sujeta a verificación humana.

Precisamente por esa incidencia sobre la discrecionalidad policial en espacio público, el caso *R (Bridges) v Chief Constable of South Wales* (2020). *Police* examinó el despliegue desde un prisma de legalidad y garantías, subrayando que su uso debe estar normativamente constreñido y evaluado en términos de proporcionalidad y salvaguardas. Así, el ámbito del sistema es el de seguridad y policía, y su etapa típica es pre-investigativa u operativa (antes o durante la intervención), donde la tecnología pasa a formar parte de la cadena de decisiones estatales que pueden traducirse en restricciones fácticas a la libertad de circulación y en prácticas de vigilancia selectiva (Davies et al., 2018; South Wales Police, 2025; R (Bridges) v Chief Constable of South Wales Police, 2020).

6.3.2. Contexto institucional, población y territorio impactados

AFR Locate impacta, en primer lugar, a los miembros del público que transitan por el área de despliegue, porque el sistema opera capturando imágenes faciales en tiempo real

y comparándolas con una watchlist definida por la autoridad policial. En otras palabras, no se limita a personas sospechosas previamente individualizadas en el terreno, sino que funciona sobre el flujo general de transeúntes dentro del campo de visión de las cámaras, y solo cuando el sistema genera una alerta (posible coincidencia) se abre la posibilidad de una intervención policial posterior. Este rasgo de captura masiva en espacio público, con activación operativa condicionada por una alerta y verificación humana aparece descrito en el propio juicio de apelación sobre AFR Locate.

En segundo lugar, el impacto se concentra en personas incluidas en la watchlist, porque son ellas quienes están materialmente expuestas a que la alerta derive en acciones policiales (p. ej., aproximación, control, verificación y, eventualmente, uso de potestades legales según el caso). Por diseño, la población "directamente buscada" es la del listado; pero por funcionamiento, la población "afectada por el escaneo" incluye a todas las personas presentes en el perímetro de despliegue.

Territorialmente, AFR Locate se implementó como una tecnología situacional, es decir, no cubre todo un territorio de forma continua, sino que se despliega en zonas específicas (normalmente de alta afluencia) conforme a objetivos operativos. En el litigio Bridges se discuten despliegues concretos en zonas geográficamente delimitadas en fechas determinadas, lo que muestra que el territorio impactado se delimita por operativos puntuales y por su justificación táctica.

6.3.3. Datos: procedencia, construcción y riesgos de sesgo

En AFR Locate, la procedencia de los datos se estructura en dos corrientes principales: (i) los datos capturados en tiempo real en el espacio público (imágenes faciales

de transeúntes dentro del perímetro de despliegue) y (ii) los datos de referencia contenidos en una watchlist previamente definida por la autoridad (imágenes e identificadores de las personas buscadas o de interés operativo). Esta arquitectura de entrada se presenta en la documentación de evaluación de impacto en protección de datos, que describe el tratamiento como procesamiento de datos personales especialmente biométricos para realizar comparaciones automatizadas con fines de identificación operativa (South Wales Police, 2025).

En cuanto a la construcción del dato biométrico, el sistema transforma las capturas faciales en una representación matemática (plantilla o firma facial) que permite ejecutar la comparación con las plantillas derivadas de la watchlist, generando así el insumo básico para una posible alerta. La evaluación independiente del uso policial enfatiza precisamente esta lógica: AFR Locate opera con watchlists relativamente acotadas (a diferencia de modalidades retrospectivas con bases masivas), y su funcionamiento depende de la extracción de características y su comparación mediante puntajes de similitud, sin que el detalle fino de la lógica propietaria sea plenamente accesible para escrutinio público (Davies et al., 2018).

Los riesgos (proxies) no se agotan en si el sistema usa o no una categoría sensible explícita, sino que aparecen por dos vías metodológicamente relevantes. Primero, por criterios de inclusión en la watchlist: aun si la watchlist se define con fines legítimos, la selección puede operar como un proxy de condiciones sociales o territoriales (quién es vigilable y por qué), amplificando patrones previos de vigilancia selectiva si no existen controles claros y trazables sobre su construcción (Fussey & Murray, 2019). Segundo, por condiciones de captura y calidad del dato: variables contextuales (ángulo, iluminación,

distancia, densidad de flujo peatonal) afectan la calidad de la imagen y, con ello, la probabilidad de coincidencias erróneas (falsos positivos); este riesgo técnico es especialmente sensible porque el sistema procesa rostros de personas no incluidas en la watchlist, de modo que la carga del error puede distribuirse de manera desigual según el lugar, el operativo y el perfil demográfico de quienes transitan por el área (Davies et al. 2018).

6.3.4. Tratamiento de datos y flujos de información

En los documentos operativos del despliegue de AFR Locate se describe un tratamiento de datos que inicia antes del uso en calle: (i) se construye y aprueba una watchlist para un objetivo policial específico y se exporta para el operativo; (ii) esa watchlist se importa en la aplicación de LFR inmediatamente antes del despliegue (y no más de 24 horas antes), y el sistema se utiliza como apoyo para localizar a personas incluidas en dicha lista; (iii) durante el despliegue, se registran métricas (p. ej., número de alertas, verdaderas o falsas, y resultados de las interacciones) y se exige mantener el uso bajo revisión cuando el operativo se prolonga, incluyendo cuestiones de retención y borrado (South Wales Police, 2025a).

En materia de conservación y eliminación, se fijaron reglas relativamente claras para el tratamiento de los datos, cuando el sistema no genera alerta, los datos biométricos del transeúnte se eliminan automáticamente de forma inmediata; cuando sí genera alerta, los datos personales asociados se eliminan tan pronto como sea practicable y, en todo caso, dentro de 24 horas (South Wales Police, 2025a).

6.3.5. Modelo o arquitectura del sistema

En términos de tipo de sistema, AFR Locate se configura como una tecnología de reconocimiento facial automatizado en tiempo real que funciona por comparación biométrica: toma rostros captados en una transmisión en vivo, construye una representación matemática y la compara contra una watchlist predeterminada, produciendo posibles coincidencias que deben ser revisadas por operadores humanos (Davies et al., 2018).

Sobre características y variables en su sentido operativo, lo públicamente reconstruible se describe a un nivel alto: el sistema detecta imágenes con rasgos de rostro, mide y analiza relaciones y distancias entre rasgos faciales, genera la firma biométrica y compara esa firma contra las imágenes almacenadas en la watchlist para estimar similitud (Davies et al., 2018).

6.3.6. Salida del sistema y alcance inferencial

La salida principal de AFR Locate se expresa como una alerta operativa de posible coincidencia cuando el sistema identifica que un rostro capturado en vivo supera el umbral técnico interno de similitud respecto de una imagen incluida en la watchlist. En la práctica, el output no se presenta como una recomendación jurídica ni como una decisión automática, sino como un disparador procedimental en donde se activa un flujo de verificación y, eventualmente, una interacción policial posterior (South Wales Police, 2025). Esta forma de salida —alerta y posterior verificación— es consistente con la descripción evaluativa del uso policial del reconocimiento facial automatizado, donde el sistema opera como mecanismo de comparación biométrica que produce candidatos a coincidencia y no identidades concluyentes (Davies et al., 2018).

En términos de umbral y categorías, lo públicamente reconstruible es, ante todo, la existencia de un umbral de activación que separa en dos categorías: sin alerta vs. alerta, más que una escala de riesgo tipo bajo, medio o alto.

Respecto de la explicación disponible, la salida es explicable solo en un sentido limitado: el afectado o los operadores pueden conocer que existió una alerta y pueden verificar elementos externos (p. ej., la imagen de referencia en la watchlist frente a la captura en vivo), pero no necesariamente pueden comprender —con trazabilidad técnica completa— por qué el algoritmo concluyó que se superó el umbral, qué rasgos fueron decisivos o cuál fue el peso relativo de la información biométrica en el match. Los documentos operativos describen la gestión de la alerta y la necesidad de verificación humana, pero no abren la lógica interna de transformación biométrica, de modo que el nivel de explicabilidad queda condicionado por el carácter propietario del sistema (South Wales Police, 2025; Davies et al., 2018).

6.3.7. Rol en la decisión estatal

En el plano formal y procedimental, la salida de AFR Locate no se configura como una decisión automática, sino como un insumo operativo que prioriza la atención policial: el sistema genera una alerta de posible coincidencia y esa alerta debe pasar por verificación humana antes de cualquier actuación (South Wales Police, 2025).

Sin embargo, el rasgo metodológicamente sensible es que, aun presentado como "apoyo", el sistema puede adquirir peso práctico en la toma de decisiones de calle en la medida en que una alerta numérica o visual se integra al flujo operativo, puede funcionar

como ancla o como señal técnica que empuja la intervención, incluso cuando jurídicamente se insista en que no determina por sí misma el curso de acción.

6.3.8. Gobernanza, evaluación de impacto y marco jurídico declarado

La gobernanza de AFR Locate (LFR) se organiza, en primer lugar, alrededor de una autoridad policial usuaria y operadora: South Wales Police despliega el sistema en operaciones públicas, define el marco interno de uso y articula el fundamento jurídico declarado para su empleo en funciones de seguridad pública (South Wales Police, 2023; South Wales Police, 2025). En esa cadena, la responsabilidad operativa inmediata recae en los equipos policiales que configuran el despliegue, gestionan la watchlist y ejecutan el flujo de verificación humana cuando se produce una alerta, conforme a reglas procedimentales internas.

En cuanto a evaluación de impacto y controles, el instrumento central de gobernanza ex ante es la Data Protection Impact Assessment (DPIA), que funciona como pieza de justificación y mitigación: identifica categorías de datos, finalidades del tratamiento, riesgos previsibles y salvaguardas (South Wales Police, 2025). A ello se suman controles de gobernanza operativa descritos en la política y procedimientos del despliegue (p. ej., condiciones para el uso, control de la watchlist, y requisitos de operación y supervisión), que buscan evitar que la herramienta opere de forma autónoma y que su uso permanezca acotado a finalidades definidas (South Wales Police, 2025).

Finalmente, el caso judicial evidencia que la gobernanza no es solo interna, sino también reactiva (ex post): R (Bridges) v Chief Constable of South Wales Police (2020) examinó si el marco normativo y de control era suficientemente delimitado y compatible con

garantías, identificando déficits en la forma en que se estructuraban ciertos márgenes de discrecionalidad y exigencias de evaluación (incluida la dimensión de igualdad), lo que obligó a reajustar el estándar de controles aplicables. En este sentido, la responsabilidad queda fragmentada: el despliegue y su justificación inmediata son de la policía, pero el estándar de legitimidad y las correcciones institucionales se consolidan por la vía de la revisión judicial y del cumplimiento de salvaguardas documentadas.

6.3.9. Transparencia, trazabilidad y auditabilidad

En AFR Locate existe un nivel de transparencia documental relevante: la autoridad policial publica instrumentos que describen el sistema y su marco de uso, por ejemplo, evaluaciones de impacto, políticas y procedimientos operativos, lo que permite reconstruir categorías generales de datos tratados, finalidades declaradas y algunas salvaguardas institucionales. En particular, estos informes identifican el carácter biométrico del tratamiento, expone riesgos y mitigaciones, y ofrece un marco institucional de explicación sobre qué hace el sistema y con qué límites pretende operar (South Wales Police, 2025). Esta publicidad, sin embargo, no equivale a trazabilidad plena del proceso algorítmico.

Desde la perspectiva de trazabilidad (auditabilidad y explicación de por qué hubo match), el punto crítico es la brecha entre lo que puede trazarse en documentos y lo que permanece inaccesible por la naturaleza propietaria del componente tecnológico. La evaluación independiente del uso por South Wales Police subraya que, aunque es posible describir el proceso general —captura, generación de firma, comparación y alerta—, el razonamiento técnico del algoritmo (parámetros, lógica exacta, y justificación interna del resultado) no es plenamente reproducible para actores externos, lo que limita la posibilidad

de explicar de forma verificable la razón del emparejamiento más allá de la comparación visual posterior (Davies et al., 2018).

Finalmente, un estándar externo de responsabilidad insiste en que la transparencia relevante para tecnologías de vigilancia no se satisface solo con publicar lineamientos generales, sino con marcos legales claros, criterios verificables de despliegue, y vías de escrutinio que reduzcan discrecionalidad y opacidad estructural. En esa línea, la crítica sostiene que, si el diseño normativo y la transparencia operativa no son suficientemente densos, el uso policial del reconocimiento facial tiende a operar con déficits de control democrático y claridad regulatoria, precisamente por el carácter intrusivo del tratamiento y por el peso práctico de las alertas en calle (Big Brother Watch, 2020).

6.3.10. Control humano significativo

El control humano significativo en AFR Locate se estructura en tres momentos: (i) supervisión ex ante, (ii) operación en tiempo real y (iii) revisión ex post, con una posibilidad explícita (al menos procedimental) de apartarse de la salida algorítmica.

En la supervisión ex ante, la intervención humana se expresa en la planificación y autorización del despliegue y, de forma especialmente relevante, en la construcción y control de la watchlist. Antes de que el sistema opere en calle, agentes responsables determinan el objetivo del operativo, el lugar y tiempo de despliegue y qué personas serán incluidas en la lista de referencia, de manera que el alcance práctico del sistema queda condicionado por decisiones institucionales previas. Este componente se enlaza con la lógica de control documentada en los instrumentos institucionales de evaluación de impacto y gobernanza del uso (South Wales Police, 2025).

En la operación, el CHS se manifiesta como un esquema *human-in-the-loop*: la alerta de posible coincidencia no se traduce automáticamente en una medida coercitiva, sino que exige una verificación por operadores entrenados antes de cualquier intervención. El procedimiento operativo establece que el personal debe revisar la alerta (incluida la comparación entre la imagen capturada y la imagen de watchlist) y decidir si se escala la actuación en terreno; esto preserva, al menos formalmente, un espacio de juicio humano y pretende evitar que el sistema funcione como identificación autónoma (South Wales Police, 2025).

En la revisión ex post, el control humano aparece en dos niveles: (a) registro y evaluación interna de despliegues (métricas de alertas, aciertos o error, y análisis de operación) y (b) mecanismos de responsabilidad externos o reactivos, en donde se incluye la revisión judicial, que han exigido que el marco de uso sea suficientemente delimitado y acompañado de salvaguardas, precisamente porque la tecnología opera en espacio público y puede afectar derechos de manera sensible (R (Bridges) v Chief Constable of South Wales Police, 2020). En este esquema, la posibilidad de apartarse se concreta en que los operadores pueden no actuar ante una alerta cuando la verificación humana no la confirma o cuando el contexto no justifica intervención, reforzando que el output no es equivalente a decisión sino a señal que debe ser evaluada críticamente por el agente (South Wales Police, 2025).

6.3.11. Evidencia de impactos: afectación y mecanismos de contención

La evidencia empírica más citada sobre el impacto operativo de AFR Locate proviene de la evaluación independiente realizada sobre despliegues de la policía de Gales del Sur. Allí se documenta que, en los eventos analizados, el sistema generó un volumen alto de

alertas de posibles coincidencias cuya proporción de falsos positivos fue significativa, especialmente en los primeros despliegues: por ejemplo, en uno de los operativos de mayor escala (final de la UEFA Champions League 2017), se registraron miles de alertas y una fracción mínima de coincidencias verdaderas. En el agregado del periodo examinado, el informe resume miles de alertas con una diferencia marcada entre coincidencias reales y falsas coincidencias, lo que revela que el impacto del sistema no se agota en su promesa de eficiencia, sino que incluye costos operativos y riesgos de intervención sobre personas no incluidas en listas de búsqueda (Davies et al., 2018).

Ahora bien, esa evidencia también muestra que el desempeño del sistema es altamente sensible al contexto (condiciones del entorno, iluminación, ángulos, calidad de cámara y configuración), lo cual explica variaciones entre despliegues: en algunos eventos el rendimiento relativo mejora, pero sin desaparecer el riesgo de falsos positivos y sin que el sistema deje de requerir verificación humana previa a cualquier actuación. El propio reporte destaca que, al degradarse condiciones como la iluminación, disminuye la capacidad de reconocimiento, lo que incrementa el margen de error y refuerza la necesidad de comprender el output como una señal probabilística y no como identificación concluyente (Davies et al., 2018).

En paralelo, la evidencia técnica general sobre reconocimiento facial aporta un punto: aun cuando un sistema no use categorías sensibles como variables explícitas, existen diferencias de desempeño por demografía (en particular, tasas de falsos positivos desiguales entre grupos), lo que puede traducirse en usos policiales en disparidades de afectación. Los reportes institucionales documenta diferencias amplias en tasas de falsos positivos entre

grupos demográficos, lo que vuelve metodológicamente pertinente tratar el sesgo como un impacto posible del sistema en condiciones reales de despliegue.

En cuanto a contención (respuestas institucionales y jurisdiccionales), el caso judicial R (Bridges) constituye el hito central: la Corte de Apelación examinó el uso de AFR Locate en despliegues concretos y concluyó que existían fallas relevantes en el marco de control, en especial por el margen de discrecionalidad sobre quién podía integrarse a una watchlist y dónde podía desplegarse el sistema; además, abordó críticamente el cumplimiento del deber de igualdad del sector público y la suficiencia de las evaluaciones y controles asociados (R (Bridges) v Chief Constable of South Wales Police, 2020).

Finalmente, también aparecen mecanismos operativos de contención descritos en el expediente judicial: (i) verificación humana de las alertas antes de actuar, (ii) criterios de gobernanza de watchlist (incluyendo límites de tamaño y umbrales de alerta como referencia), y (iii) prácticas de borrado de imágenes de no coincidencias, que buscan reducir retención innecesaria. No obstante, el propio análisis judicial subraya que estas medidas no sustituyen un estándar robusto de legalidad, previsibilidad y control, especialmente cuando la tecnología produce errores con potencial de impacto diferenciado (R (Bridges) v Chief Constable of South Wales Police, 2020).

6.3.12. Síntesis del caso

Una vez realizada la reconstrucción técnico institucional del caso, a continuación se presenta a manera de resumen los hallazgos de la caracteizacion técnico-institucional de forma estructura como se había establecido previamente en la metodología Tabla 1 cuya función es condensar en un formato homogéneo los elementos relevantes del sistema

analizado (finalidad, población impactada, datos, tratamiento, modelo, salida, rol, gobernanza, transparencia/trazabilidad, control humano significativo y evidencia de impactos), permitiendo visualizar de manera integrada cómo se articula el dispositivo y dónde se ubican los principales puntos de tensión.

Tabla 6
Síntesis de la caracterización técnico-institucional del caso AFR Locate

Dimensión	Preguntas guía	Hallazgos clave
Finalidad declarada y contexto	¿Qué problema pretende resolver? ¿En qué decisión o incidencia estatal se inserta?	AFR Locate (reconocimiento facial en vivo) se presentó como una herramienta preventiva e investigativa para identificar, en tiempo real y en espacios públicos, a personas incluidas en una watchlist. Buscó apoyar la seguridad pública y la gestión operativa en despliegues (eventos, zonas concurridas), habilitando verificaciones de identidad, detenciones o acciones de salvaguarda tras confirmación humana. Su uso incidió en decisiones policiales inmediatas.
Población y territorio impactado	¿A quién aplica? ¿Se focaliza territorial o poblacionalmente?	El impacto directo recae en quienes están en la watchlist y en el público general que transita por el área de despliegue, cuyas imágenes se capturan y comparan en vivo. Territorialmente, se despliega en ubicaciones específicas dentro de la jurisdicción policial (centros urbanos, estadios, eventos), según autorizaciones operativas y criterios de necesidad y proporcionalidad. La inclusión en watchlists y la selección de lugares exigen reglas claras para evitar discrecionalidad excesiva.
Datos: origen y calidad	¿De dónde vienen los datos? ¿Son pertinentes, representativos, actualizados?	El sistema usa (i) imágenes capturadas en vivo y (ii) imágenes de referencia de la watchlist. De allí genera plantillas biométricas y calcula similitud. Los

Dimensión	Preguntas guía	Hallazgos clave
		riesgos surgen menos por variables sensibles explícitas y más por condiciones de captura, composición de la watchlist y diferencias de desempeño por demografía reportadas en pruebas técnicas, lo que vuelve central evaluar impactos en igualdad y no discriminación
Tratamiento de datos	¿Cómo se recolectan, cruzan, conservan y comparten?	Flujo: captura → detección → plantilla → comparación → puntaje → alerta (si supera umbral); después verificación humana y, si procede, actuación. En cuanto a su conservación, las imágenes sin alerta se borran de inmediato; las asociadas a alertas se retienen solo para revisión y se eliminan en corto plazo. La watchlist del despliegue se suprime tras su uso, aunque algunos detalles de retención e intercambio no son verificables públicamente.
Modelo y arquitectura	¿Qué tipo de modelo (reglas/ML/DL)? ¿Qué variables usa?	AFR Locate convierte rostros en representaciones biométricas y calcula un puntaje de similitud frente a un conjunto cerrado (watchlist). Los documentos públicos describen la lógica general (captura, plantilla, comparación, puntaje y alerta), pero no permiten precisar el algoritmo del proveedor, su entrenamiento o parámetros internos: esa dimensión opera como componente propietario (caja negra), relevante para gobernanza y auditoría. Los ajustes visibles suelen ser operativos (umbrales y parámetros de despliegue).
Salida del sistema	¿Qué produce (score, alerta, recomendación)?	El output es una alerta de posible coincidencia cuando el puntaje supera el umbral, con imagen capturada y de referencia para verificación. No implica identificación automática, exige confirmación de un operador antes de intervenir. Su alcance es de priorización y soporte, no de certeza

Dimensión	Preguntas guía	Hallazgos clave
		individual. La confiabilidad depende del entorno de captura, parámetros técnicos y calidad de la watchlist, por lo que el margen de error debe asumirse como un dato relevante del uso.
Rol en la decisión	¿La salida decide, recomienda o prioriza?	Formalmente, el sistema sugiere una coincidencia y prioriza, pero la decisión estatal (detener, identificar, intervenir) la adopta el agente tras verificación humana. El riesgo es la automatización de facto: en contextos de presión, la alerta puede anclar o reforzar la decisión, aunque sea solo un indicio. Por eso el control judicial insiste en reglas claras, límites a la discrecionalidad y evaluaciones de impacto adecuadas.
Gobernanza, evaluación de impacto y responsabilidades	¿Quién opera? ¿Quién responde? ¿Qué controles y evaluaciones existen?	La gobernanza combina un operador público y un proveedor comercial. Documentos institucionales fijan mandato legal, política de despliegue y evaluación de impacto. La responsabilidad por autorizaciones y actuaciones recae en la autoridad policial, no en el proveedor. Los controles ex ante se expresan en DPIA/políticas y los ex post se activan por revisión judicial, que identifica déficits y obliga a ajustar salvaguardas.
Transparencia y trazabilidad	¿Es auditable? ¿Hay explicabilidad o registro?	Existe transparencia documental sobre finalidades, flujos de datos, roles y mitigaciones. Pero la trazabilidad de por qué se produce un match es limitada: se ve el resultado (alerta) y la verificación humana, no el razonamiento interno del algoritmo. Críticas externas advierten que publicar documentos no sustituye un marco legal claro ni rendición de cuentas robusta cuando se afectan derechos.

Dimensión	Preguntas guía	Hallazgos clave
Control humano significativo (CHS)	¿Dónde interviene el humano (diseño-uso-revisión)?	El CHS aparece (i) ex ante, en la autorización del despliegue y la gobernanza de watchlists; (ii) durante la operación, en la revisión obligatoria de cada alerta y la posibilidad real de descartarla; y (iii) ex post, en registros, revisión de incidentes y control judicial. La clave es que el humano valida y puede apartarse del sistema, aunque persiste el riesgo de sesgo de automatización.
Evidencia de impactos	¿Hay evidencia de sesgos/errores/afectaciones?	La evidencia empírica reporta falsos positivos relevantes: en despliegues evaluados, la mayoría de alertas fue descartada como falsa coincidencia, con riesgo de intervenciones erróneas y carga operativa. En lo jurídico, Bridges señaló déficits de legalidad y del deber de igualdad, impulsando revisiones de salvaguardas. Pruebas técnicas amplias documentan diferencias de desempeño por demografía, lo que exige monitoreo y mitigación continuos.

6.4 Caso PRISMA (Colombia)

6.4.1. Finalidad declarada y decisión estatal afectada

PRiSMA (*Perfil de Riesgo de Reincidencia para la Solicitud de Medidas de Aseguramiento*) fue una herramienta desarrollada institucionalmente por la Fiscalía General de la Nación a través de su Dirección de Políticas y Estrategia como parte de una agenda de política criminal basada en evidencia. Su finalidad declarada parte de dos justificaciones complementarias: (i) eficiencia, orientada a reducir reincidencia y hacer un uso más razonable de los cupos carcelarios escasos; y (ii) justicia y proporcionalidad, dirigida a concentrar la medida intramural en quienes presenten niveles objetivos altos de riesgo de

reincidencia, y a disminuir errores en la solicitud y otorgamiento de dichas medidas (Fiscalía General de la Nación, 2019).

En términos de decisión estatal afectada, PRiSMA se vincula directamente con una de las determinaciones más intensas del proceso penal: si se solicita y eventualmente se impone la detención preventiva intramural como medida de aseguramiento. En la reconstrucción institucional, la herramienta se concibe para complementar la información con la que cuentan los fiscales en audiencias de solicitud de medida de aseguramiento, de modo que el perfil de riesgo opere como insumo para sustentar o racionalizar la petición de privación de libertad antes de condena (Fiscalía General de la Nación, 2019).

Justamente por insertarse en medidas cautelares privativas de libertad, su propósito declarado (racionalización y eficiencia) queda metodológicamente atado a un punto de fricción garantista: el riesgo de que una predicción algorítmica reconfigure, en la práctica, el estándar de decisión sobre restricción de libertad y las garantías asociadas. En ese sentido, el reporte regional de Fair Trials (2024) ubica a PRiSMA como un sistema de evaluación de riesgo de reincidencia orientado a apoyar la decisión de la Fiscalía sobre solicitar o no detención preventiva, y advierte preocupaciones por su potencial incidencia en garantías como defensa, presunción de inocencia e imparcialidad judicial.

6.4.2. Contexto institucional, población y territorio impactados

La herramienta PRiSMA impacta, ante todo, a personas imputadas o acusadas dentro del proceso penal colombiano, específicamente en el momento en que la Fiscalía General de la Nación debe decidir si solicita o no una medida de aseguramiento (detención preventiva u

otras medidas), de modo que su alcance real se proyecta sobre decisiones estatales que distribuyen cargas de restricción (privación preventiva de la libertad) y beneficios de permanencia en libertad durante la investigación. En términos de escala poblacional, la reconstrucción disponible sugiere que PRiSMA se alimenta de una base con información masiva de millones de individuos con registros penales, lo cual indica que el universo potencialmente afectado no es marginal, sino estructural dentro del funcionamiento del sistema penal (Fair Trials, 2024).

Desde el punto de vista territorial e institucional, la evidencia pública más concreta sobre despliegue se ubica en el plan piloto: un piloto con 10 fiscales URI en cinco direcciones seccionales (Ibagué, Cartagena, Bogotá, Popayán y Medellín), donde los fiscales descargan en tiempo real desde el SPOA un documento digital (PDF) con el resumen del riesgo e historial para usarlo en la audiencia de solicitud de medida de aseguramiento (Fiscalía General de la Nación, 2019). En consecuencia, el impacto territorial verificable se concentra, al menos en esta evidencia, en esas seccionales piloto y en los flujos de decisión asociados a las URI, sin que los documentos disponibles permitan afirmar con el mismo grado de certeza un despliegue nacional uniforme más allá de ese piloto.

6.4.3. Datos: procedencia, construcción y riesgos

Los documentos públicos disponibles indican que PRISMA se alimenta de registros administrativos y judiciales integrados por Fiscalía General de la Nación, con cruces entre SPOA, SIEDCO e información del INPEC para reconstruir trayectorias penales y penitenciarias relevantes para la predicción (Fiscalía General de la Nación, s. f.).

En cuanto a cómo se construye el dato, es decir, qué se considera relevante, la reconstrucción pública presenta los insumos en bloques de información asociados al individuo y al evento, y a historiales previos: delitos previos, capturas previas, medidas previas e información penitenciaria, usados como componentes del modelo. Además, en el diseño reportado por la Fiscalía, la variable de respuesta (lo que se toma como resultado a predecir) se construye mediante seguimiento del caso para identificar un evento posterior (p. ej., una nueva medida o condena), con soporte en las bases mencionadas (Fiscalía General de la Nación).

El punto crítico es la calidad y el riesgo de proxies: si los insumos provienen de registros policiales y de persecución penal, parte del dato puede incorporar huellas institucionales (selectividad de vigilancia, prácticas de captura e imputación, desigualdades territoriales), lo que vuelve plausible que variables aparentemente neutrales funcionen como proxies de condiciones socioeconómicas o territoriales. En términos de riesgo, el mismo reporte regional advierte que, cuando decisiones judiciales se apoyan en sistemas entrenados con datos de fiscalía o policía, pueden reproducirse sesgos de origen institucional y comprometer garantías asociadas a imparcialidad y control (Fair Trials, 2024).

Ahora bien, con los documentos consultados, no es posible precisar con rigor un diccionario completo de variables, reglas de depuración, tasas de faltantes, ni si se excluyen o cómo se tratan variables sensibles directas.

6.4.4. Tratamiento de datos y flujos de información

El tratamiento de datos en PRISMA, según lo que es reconstruible a partir de documentos públicos, inicia con la recolección por extracción de información administrativa y penal ya existente en los sistemas estatales, en particular desde el SPOA y el SIEDCO, y con la integración de información penitenciaria asociada al INPEC, con el propósito de consolidar un perfil individual que permita estimar riesgo de reincidencia en el contexto de la solicitud de medida de aseguramiento. Esa recolección no corresponde a un formulario nuevo diligenciado por el evaluado, sino al aprovechamiento de registros institucionales acumulados, descritos como una base histórica disponible a nivel individual desde 2005 y con millones de registros asociados a antecedentes y eventos dentro del sistema penal, lo que condiciona el tratamiento hacia el cruce y depuración de grandes volúmenes de datos administrativos (Fiscalía General de la Nación, 2019).

Respecto de la conservación y retención de los datos, la evidencia pública permite afirmar que el tratamiento se apoya en una base histórica amplia y que el sistema opera sobre registros acumulados en el tiempo, pero no es posible precisar, con los documentos disponibles, políticas concretas de retención, plazos de conservación, reglas de borrado, ni protocolos detallados de acceso, auditoría interna o parametros asociados al ciclo completo del dato. En el mismo sentido no se encuentra suficientemente documentado en las fuentes si el reporte que el sistema entrega al fiscal se incorpora formalmente como pieza del expediente, si se entrega de forma sistemática a la defensa, ni si existen terceros tecnológicos con acceso a los datos o a la infraestructura por ejemplo, proveedores de analítica o alojamiento,

6.4.5. Salida del sistema y alcance inferencial

La salida principal del sistema PRiSMA se materializa en un reporte descargable en formato PDF, pensado para complementar la información disponible en las audiencias de solicitud de medida de aseguramiento, y que puede generarse en tiempo real a partir de la información integrada en los sistemas institucionales. En su estructura, el reporte presenta como resultado central una estimación de probabilidad de reincidencia, derivada de un modelo de aprendizaje supervisado que predice el riesgo dadas las características del individuo, del último evento criminal y de los antecedentes criminales. Ese resultado no se limita a un único indicador, la presentación institucional contempla probabilidades por tipología, incluyendo reincidencia general, crimen a la propiedad, crimen violento y otros delitos (Fiscalía General de la Nación, s. f.).

En cuanto a forma de clasificación, el reporte ejemplificado muestra porcentajes de riesgo acompañados de una escala cualitativa (p. ej., bajo/medio/alto) representada visualmente mediante barras, lo que facilita su lectura operativa en el marco de decisiones rápidas. A la par, el output incorpora un perfil informativo que contextualiza el resultado, secciones de evento actual, resúmenes de actuaciones y antecedentes construidos con información procedente del SPOA y de fuentes administrativas y penales integradas incluyendo tablas y trazas cronológicas del historial disponible (Fiscalía General de la Nación, s. f.).

Finalmente, sobre el alcance inferencial, PRiSMA define su output como probabilidad estimada (no como constatación), y en la evaluación institucional se ilustra que el riesgo se distribuye en un continuo (0 a 1), con contrastes entre extremos como el 10% menos riesgoso frente al 10% más riesgoso (Fiscalía General de la Nación, s. f.,). En coherencia con esa lógica, la herramienta se presenta como un insumo para orientar

priorización bajo el criterio de concentrar medidas intramurales en quienes exhiben altos niveles objetivos de riesgo, lo cual refuerza que su salida pretende funcionar como indicador probabilístico para apoyar decisiones, más que como determinación automática de la medida aplicable (Fiscalía General de la Nación).

6.4.6. Rol en la decisión estatal

En su configuración declarada, PRiSMA no aparece diseñado para decidir automáticamente la imposición de una medida de aseguramiento, sino para apoyar la decisión del fiscal mediante un insumo probabilístico que acompaña la preparación de la solicitud en audiencia. A la vez, diseño presenta umbrales y segmentaciones, por ejemplo, el uso de cortes de riesgo para identificar perfiles más riesgosos muestra que la salida funciona también como mecanismo de priorización, permite ordenar casos por nivel de riesgo y orientar el foco institucional hacia un subconjunto de imputados según el puntaje del modelo (Fiscalía General de la Nación).

Ahora bien, en términos de automatización de facto, el riesgo metodológicamente relevante no depende de que el sistema decida formalmente, sino de que el puntaje se convierta en un ancla práctica para la actuación fiscal y, por esa vía, module la dinámica de la audiencia.

Con base en los documentos analizados no es posible reconstruir con precisión (i) si existe un protocolo formal de apartamiento documentado frente al puntaje, (ii) si el reporte se entrega sistemáticamente a la defensa como insumo controvertible en audiencia, o (iii) qué reglas internas delimitan el peso del puntaje en la decisión de solicitar la medida.

6.4.7. Gobernanza institucional y marco jurídico declarado

En términos operativos, el esquema sugiere una gobernanza intraestatal: (i) la Fiscalía define el propósito, integra fuentes y pone la herramienta al servicio de su función acusatoria, y (ii) los funcionarios que tramitan el caso utilizan el reporte como insumo para sustentar la petición ante el juez (Fiscalía General de la Nación).

En cuanto al marco jurídico declarado, la propia documentación de PRiSMA lo ubica en el contexto de la medida de aseguramiento del proceso penal y la necesidad de sustentarla con elementos que permitan inferir ciertos riesgos procesales o materiales. De forma concordante, la regulación procesal establece los requisitos y condiciones para la procedencia de la medida de aseguramiento y delimita el rol de la Fiscalía como solicitante frente al juez de control de garantías como decisor. En esta clave, la responsabilidad decisional (privación o restricción de libertad) no se traslada formalmente al sistema, el output se incorpora como soporte argumentativo dentro de un trámite judicial que, en principio, exige motivación y contradicción.

En la práctica, la gobernanza que puede reconstruirse desde fuentes públicas muestra controles parciales y un componente de pilotaje como forma de gestión del riesgo: la evidencia disponible indica que PRiSMA pasó por pruebas piloto orientadas a identificar errores y posibles sesgos antes o durante su adopción. Sin embargo, no es verificable públicamente un esquema integral y estable de evaluación de impacto en derechos, auditoría externa periódica, o reglas completas de rendición de cuentas sobre funcionamiento y actualización del sistema; esa insuficiencia documental es, en sí misma, un hallazgo de gobernanza.

6.4.8. Transparencia, trazabilidad y auditabilidad

En términos de transparencia, la información públicamente accesible sobre PRiSMA se concentra, sobre todo, en una presentación institucional de la Fiscalía General de la Nación, donde se enuncia (i) la finalidad de eficiencia y justicia en el uso de medidas de aseguramiento y (ii) una descripción del esquema de predicción: aprendizaje supervisado y el tipo de técnica empleada, junto con categorías generales de variables (características del individuo, del evento, delitos, capturas, medidas previas e información penitenciaria del INPEC).

Ahora bien, esa transparencia es insuficiente para la auditabilidad plena puesto que en las fuentes públicas consultadas no se publica el código, ni un detalle verificable de (i) la lista completa de variables y sus transformaciones, (ii) los pesos o estructuras internas del modelo, (iii) protocolos de validación externa replicable (por terceros) y sus métricas completas, o (iv) umbrales operativos que conecten puntajes con reglas internas de uso institucional.

En el mismo sentido, reportes regionales señalan que PRiSMA se usó para apoyar la decisión de la Fiscalía sobre si solicitar o no detención preventiva, pero también subrayan preocupaciones por garantías procesales y que no se tiene información de que el sistema continúe utilizándose, lo cual es indicativo de un entorno de información pública limitada sobre el ciclo de vida y controles del sistema (Fair Trials, 2024).

En trazabilidad y registro, el punto crítico no es tanto saber que existen insumos y un output, sino poder seguir la ruta del dato y del resultado: quién accede, cómo se registra la consulta, cómo se conserva el reporte, por qué y con qué versión del modelo se generó un

perfil. Esos extremos (retención, interoperabilidad detallada, protocolos de acceso, y trazabilidad interna, y mecanismos de explicación individual al afectado) no aparecen desarrollados en la documentación pública revisada; porque esa falta de trazabilidad pública es en sí misma un hallazgo de gobernanza relevante para las garantías (Universidad de los Andes, 2024).

6.4.9. Control humano significativo

En el uso operativo, el control humano significativo se concreta en que el fiscal decide si genera y utiliza el reporte, y en que ese reporte entra al trámite como insumo dentro de la construcción argumentativa de la solicitud. Aun cuando la herramienta priorice o recomiende mediante un puntaje, la decisión estatal formalmente determinante sobre privación preventiva de la libertad permanece en el juez de control de garantías, de acuerdo con la estructura procesal de la medida de aseguramiento. Esto permite sostener, en clave CHS, que existe (i) intervención humana decisoria en el ente acusador (usar o no usar; cómo argumentar) y (ii) intervención humana decisoria final en sede judicial (imponer o no imponer), lo cual ubica a PRiSMA en un esquema de apoyo y no de automatización total.

En la revisión ex post y la posibilidad real de apartarse, la evidencia pública es significativamente más débil. En los documentos consultados no se describen con precisión (i) protocolos de apartamiento documentado del puntaje por razones motivadas, (ii) exigencias internas de justificación cuando el fiscal decide contrariar el puntaje, (iii) auditorías periódicas que revisen sesgos, error o cambios por subpoblación, ni (iv) mecanismos sistemáticos de trazabilidad que permitan evaluar, a posteriori, cuánto pesó el

output en solicitudes y decisiones judiciales. En ese marco, el CHS puede afirmarse desde una lógica formal, más no es posible garantizarla desde un punto de vista material.

6.4.10. Evidencia de impactos: afectaciones y mecanismos de contención

En la documentación institucional, el principal impacto atribuido a PRiSMA se formula en términos de eficiencia y gestión del riesgo en la detención preventiva: si se mantiene constante el número anual de medidas intramurales (aprox. 24.000 en 2018), el uso del modelo permitiría reducir en 25% los delitos cometidos por reincidentes; o, alternativamente, si se mantiene constante el nivel de delitos cometidos por reincidentes, permitiría reducir en 36% el número de medidas intramurales (Fiscalía General de la Nación, s. f.).

En sentido crítico, la evidencia secundaria disponible en la región que manifiesta pese a promesas de garantías, no es posible verificar públicamente un balance robusto de eficiencia, ni impactos materia de igualdad y no discriminación, ni una evaluación abierta de efectos sobre derechos, y señala que no hay información pública suficiente para afirmar que PRISMA continúe en uso (Fair Trials, 2024). En una línea convergente, una ficha del observatorio académico de la Universidad de los Andes (2025) sobre algoritmos y derechos reporta que PRISMA se encuentra suspendido, lo que, como hecho metodológicamente relevante, desplaza el eje de impacto desde la eficacia hacia las insuficiencias de garantías y control que rodearon su despliegue.

6.4.11. Síntesis del caso

Una vez realizada la reconstrucción técnico institucional del caso, a continuación se presenta a manera de resumen los hallazgos de la caracterización técnico-institucional de forma estructura como se había establecido previamente en la metodología Tabla 1 cuya función es condensar en un formato homogéneo los elementos relevantes del sistema analizado (finalidad, población impactada, datos, tratamiento, modelo, salida, rol, gobernanza, transparencia/trazabilidad, control humano significativo y evidencia de impactos), permitiendo visualizar de manera integrada cómo se articula el dispositivo y dónde se ubican los principales puntos de tensión.

Tabla 7

Síntesis de la caracterización técnico-institucional del caso PRiSMA

Dimensión	Preguntas guía	Hallazgos clave
Finalidad declarada y contexto	¿Qué problema pretende resolver? ¿En qué decisión/incidencia estatal se inserta?	PRISMA es una herramienta de Fiscalía General de la Nación para apoyar al fiscal al sustentar la solicitud de medida de aseguramiento (intramural, domiciliaria o ninguna), mediante una estimación del riesgo de reincidencia. Su finalidad declarada es mejorar consistencia y eficiencia en decisiones que distribuyen cargas de restricción de libertad y supervisión, aportando información para audiencias y escritos, sin reemplazar formalmente el juicio del operador.
Población y territorio impactado	¿A quién aplica? ¿Se focaliza territorial o poblacionalmente?	Impacta a personas adultas imputadas o investigadas en el proceso penal colombiano, especialmente cuando se evalúa detención preventiva.
Datos: origen y calidad	¿De dónde vienen los datos? ¿Son pertinentes, representativos, actualizados?	El sistema se alimenta de fuentes administrativas: SPOA y otros registros del sistema penal, más información de capturas, antecedentes y datos penitenciarios. Aunque se afirma ciego a variables sensibles, varias entradas (historial de capturas, territorialidad, educación, etc) pueden operar como proxies de estatus socioeconómico y de sesgos históricos de policiamiento; la calidad del dato depende de cómo se produce y actualiza el registro.
Tratamiento de datos	¿Cómo se recolectan, cruzan, conservan y comparten?	El tratamiento integra recolección y cruce de bases (SPOA, registros policiales y penitenciarios) para construir variables sobre persona, hecho y trayectoria. Los reportes se generan en el ecosistema institucional y se incorporan como insumo del caso.

Dimensión	Preguntas guía	Hallazgos clave
		Sin embargo, hay información pública limitada sobre retención, acceso interno e interoperabilidad detallada, por lo que estos flujos se registran como no verificables públicamente.
Modelo y arquitectura	¿Qué tipo de modelo (reglas/ML/DL)? ¿Qué variables usa?	PRISMA se presenta como un modelo de aprendizaje supervisado; la Fiscalía describió su uso con base en variables agrupadas en características del individuo, del último hecho, historial delictivo, capturas previas, medidas previas e información INPEC. La validación se reporta de forma general, pero la especificación completa del modelo y sus ponderaciones no es plenamente pública, introduciendo un componente de "caja negra" institucional.
Salida del sistema	¿Qué produce (score, alerta, recomendación)?	La salida se entrega como un reporte que resume nivel de riesgo y orientaciones para el fiscal, pensado para apoyar si se solicita o no la medida y cuál tipo. El resultado es probabilístico y debe leerse como estimación basada en patrones previos, no como certeza individual. Umbrales de corte pueden priorizar casos, afectando la intensidad de intervención estatal.
Rol en la decisión	¿La salida decide, recomienda o prioriza?	Formalmente, el output funciona como soporte: orienta la argumentación del fiscal sobre necesidad y proporcionalidad, pero no decide por sí mismo. En la práctica, existe riesgo de automatización de facto: el puntaje puede anclar la solicitud fiscal y, por arrastre, influir en la decisión judicial cuando se percibe como "prueba técnica". Por eso, su rol se entiende como recomendación y priorización con potencial efecto determinante.
Gobernanza, evaluación de impacto y responsabilidades	¿Quién opera? ¿Quién responde? ¿Qué controles y evaluaciones existen?	La gobernanza combina desarrollo y operación intraestatal: la Fiscalía gestiona la herramienta y los fiscales la usan en su flujo; otras entidades aportan datos (INPEC y Policía). La rendición de cuentas se reparte entre equipos técnicos y operadores que incorporan el reporte al expediente. La evidencia pública sugiere una gobernanza más reactiva (pilotos y ajustes) que un esquema previo robusto de evaluación de impacto y control externo continuo.
Transparencia y trazabilidad	¿Es auditable? ¿Hay explicabilidad y registro?	La transparencia es parcial: hay documentos que describen finalidad, tipo de modelo y fuentes de datos, pero no se divulgan integralmente datasets, código, ponderaciones ni auditorías independientes completas. La trazabilidad se apoya en el reporte incorporado al expediente (qué salió y cuándo), pero la explicabilidad para el afectado es baja: es difícil reconstruir cómo cada variable incidió en su puntaje y qué cambios sufrió el modelo en el tiempo.

Dimensión	Preguntas guía	Hallazgos clave
Control humano significativo (CHS)	¿Dónde interviene el humano (diseño-uso-revisión)?	El control humano se ubica en: (i) diseño y ajustes por equipos técnicos de lo cual no es publica la información, (ii) uso operativo por el fiscal que puede complementar o controvertir el reporte con el expediente, y (iii) revisión judicial al decidir la medida. Para que sea significativo, debe existir posibilidad real de apartarse del puntaje, comprender sus límites y corregir errores. Si el reporte se toma como estándar, el control tiende a volverse meramente formal.
Evidencia de impactos	¿Hay evidencia de sesgos, errores o afectaciones?	Sobre impactos, la documentación institucional plantea beneficios potenciales (focalizar medidas y reducir riesgos), pero organizaciones y repositorios académicos señalan preocupaciones por sesgos, opacidad y debido proceso en decisiones de detención preventiva. Se ha indicado que el sistema está suspendido, lo que sugiere controversia y/o déficit de garantías. La evidencia disponible es más descriptiva y de debate público que de evaluaciones independientes concluyentes sobre efectos en derechos.

6.5 Caso Fiscal Watson (Colombia)

6.5.1. Finalidad declarada y contexto

Fiscal Watson es una herramienta de analítica de contenidos y búsqueda avanzada implementada por la Fiscalía General de la Nación, basada en tecnología IBM Watson, orientada a apoyar la investigación penal mediante el hallazgo y asociación de información dentro del Sistema Penal Oral Acusatorio (Palacios et al., 2024). En su formulación institucional, el sistema se presenta como un dispositivo para procesar grandes volúmenes de registros y textos del SPOA con el fin de identificar relaciones entre noticias criminales (casos) y producir insumos útiles para la gestión investigativa y la priorización operativa, sin desplazar formalmente el juicio del fiscal o del investigador humano (Fiscalía General de la Nación, 2019).

El contexto declarado del despliegue es el alto volumen acumulado de información en el SPOA —del orden de decenas de millones de registros— y la necesidad de dotar a los equipos de analistas de capacidades de búsqueda y analítica que permitan acelerar la asociación de casos y la detección de patrones relevantes para la intervención temprana (Fiscalía General de la Nación, 2020).

6.5.2. Contexto institucional, población y territorio impactados

El impacto de Fiscal Watson recae, de manera primaria, sobre las personas vinculadas a las noticias criminales registradas en el SPOA —víctimas, indiciados, imputados u otras personas mencionadas en los relatos y datos de los casos— en la medida en que sus datos pueden ser indexados, correlacionados y recuperados para producir hipótesis de conexión entre investigaciones (Fiscalía General de la Nación, 2019). En términos territoriales, el sistema se proyecta para operar sobre información registrada por seccionales y dependencias distribuidas en el territorio nacional, con capacidad de asociar eventos por variables espaciales y temporales, de forma que su alcance no es local sino potencialmente inter-jurisdiccional (Fiscalía General de la Nación, 2019).

Metodológicamente, su población objetivo no es un grupo demográfico delimitado, sino el universo de casos y sujetos que ingresan al SPOA, por lo que la intensidad del impacto depende de la fase procesal, la dependencia usuaria y el tipo de consulta o analítica aplicada en cada operación (Palacios et al., 2024, p. 10).

6.5.3. Datos: procedencia, construcción y riesgos de sesgo

Las fuentes de datos de Fiscal Watson se anclan principalmente en el SPOA, que concentra información estructurada y no estructurada derivada de denuncias, informes y actuaciones procesales, incluyendo relatos textuales y campos administrativos susceptibles de indexación y análisis (Palacios et al., 2024,). La documentación institucional describe que, para fines de asociación y análisis, el sistema emplea variables asociadas a hechos delictivos como georreferenciación del delito, lugares y fechas, patrones o modus operandi y variables relacionadas con reincidencia criminal, lo que permite construir vínculos entre casos con similitudes relevantes (Fiscalía General de la Nación, 2019).

En términos de calidad, la robustez del insumo depende de la completitud y consistencia de los registros del SPOA, por lo que errores de captura, omisiones o diferencias de estandarización entre dependencias pueden afectar la pertinencia de las asociaciones sugeridas. Asimismo, varias de las variables operativas (territorio, historial institucional, redes de interacción) pueden funcionar como proxies de condiciones socioeconómicas o de selectividad policial, de modo que el riesgo de sesgo no proviene de introducir categorías sensibles explícitas, sino de correlaciones estructurales inherentes a la producción del dato administrativo penal (Palacios et al., 2024).

6.5.4. Tratamiento de datos y flujos de información

El tratamiento inicia con la recolección y consolidación de información en el SPOA, desde donde Fiscal Watson extrae e indexa contenidos para habilitar búsquedas complejas y análisis de relaciones entre noticias criminales, lo cual supone el cruce de registros por coincidencias textuales, temporales y espaciales (Fiscalía General de la Nación, 2019). Los documentos públicos disponibles no permiten reconstruir con alta precisión los parámetros

de retención, políticas de conservación, protocolos de acceso granular, ni los mecanismos de intercambio inter-agencial de los repositorios indexados. Esta indeterminación documental es relevante en términos de gobernanza, porque el flujo efectivo de datos —quién consulta, qué se registra como traza y bajo qué controles— condiciona el impacto real del sistema aun cuando su finalidad declarada sea apoyo (Palacios et al., 2024).

6.5.5. Modelo o arquitectura del sistema

Fiscal Watson se configura como un sistema de analítica de información orientado a recuperación y descubrimiento de conocimiento, en el que la capacidad central no es predecir conductas futuras sino extraer y correlacionar información de grandes corpus documentales para sugerir conexiones entre casos y elementos relevantes para el analista (Palacios et al., 2024). El componente tecnológico identificado en fuentes institucionales y judiciales corresponde a IBM Watson Explorer/Content Analytics, cuya operación típica se basa en indexación de contenido, extracción de entidades y consultas avanzadas para recuperar información a partir de grandes volúmenes de texto y metadatos (Fiscalía General de la Nación, 2019). No obstante, la especificación detallada de los algoritmos de asociación (por ejemplo, ponderaciones de similitud, reglas internas y ajustes de relevancia) no es plenamente auditable desde documentos públicos y se encuentra mediada por componentes propietarios, lo que delimita el grado de explicabilidad externa del funcionamiento (Palacios et al., 2024).

6.5.6. Salida del sistema y alcance inferencial

La salida de Fiscal Watson se expresa, principalmente, como resultados de búsqueda y reportes de asociación de noticias criminales dentro del SPOA, en los que el sistema

identifica registros potencialmente relacionados y permite al analista explorar vínculos por variables espaciales, temporales y de modus operandi. De acuerdo con reportes institucionales, la herramienta produce un índice de información de las noticias criminales asociadas, destinado a orientar la asignación de recursos y la distribución de capacidades investigativas hacia concentraciones o patrones detectados en los datos (Fiscalía General de la Nación, 2019).

En términos inferenciales, estos resultados deben entenderse como hipótesis de relación derivadas de patrones en el corpus (relevancia y compatibilidad), y no como determinaciones fácticas o conclusiones probatorias autosuficientes sobre autoría o responsabilidad penal (Palacios et al., 2024). Por ello, el alcance del output es instrumental y depende de la verificación posterior por parte de operadores humanos, quienes deben contrastar las asociaciones sugeridas con evidencia y contexto del caso concreto (Fiscalía General de la Nación, 2020).

6.5.7. Rol en la decisión estatal

En el plano formal, Fiscal Watson se presenta como un insumo para apoyar la toma de decisiones de carácter operativo-administrativo y de gestión investigativa, especialmente en fases de análisis, priorización y asignación de recursos, sin sustituir la decisión discrecional del fiscal o de la autoridad competente (Fiscalía General de la Nación, 2019). En la práctica institucional descrita, el sistema se integra a equipos de analistas que utilizan la herramienta para asociar casos y producir insumos para intervención temprana, lo que puede incidir en qué investigaciones se conectan, cuáles se priorizan y qué hipótesis reciben mayor atención investigativa (Fiscalía General de la Nación, 2020).

Este rol consultivo puede convertirse en automatización de facto cuando la salida se naturaliza como criterio predominante de priorización o cuando el peso de la analítica se impone sobre otras fuentes de información por razones de eficiencia organizacional. En consecuencia, el rol debe describirse como recomendación o soporte con capacidad de orientar decisiones estatales relevantes, cuya intensidad depende del grado de dependencia institucional y de los controles de uso establecidos (Palacios et al., 2024).

6.5.8. Gobernanza institucional y marco jurídico declarado

La gobernanza de Fiscal Watson se configura como un arreglo socio-técnico en el que confluyen, por un lado, el proveedor tecnológico (IBM Watson y sus componentes de analítica) y, por otro, la Fiscalía como entidad operadora que define finalidades, escenarios de uso y responsables internos de la explotación de la herramienta (Palacios et al., 2024). En reportes institucionales se ubica la herramienta en el marco de estrategias de big data e innovación para la investigación penal, con implementación y operación dentro de dependencias encargadas del análisis y de la asociación de casos para intervención temprana (Fiscalía General de la Nación, 2020).

En materia de evaluación de impacto, la evidencia pública disponible no permite verificar la existencia de evaluaciones integrales ex ante en derechos (por ejemplo, con enfoque de igualdad o no discriminación y debida diligencia en datos), más allá de reportes de gestión y resultados agregados (Palacios et al., 2024). Por ello, la atribución de responsabilidades se distribuye entre quien diseña y actualiza componentes técnicos, quien administra y controla el acceso a los datos y quien toma decisiones sobre el uso del output, lo que exige describir la gobernanza como fragmentada y potencialmente reactiva.

6.5.9. Transparencia, trazabilidad y auditabilidad

La transparencia de Fiscal Watson se ve limitada por dos factores convergentes: la naturaleza propietaria de los componentes tecnológicos y la escasez de documentación pública detallada sobre el funcionamiento, los criterios de asociación y los protocolos de operación y control (Palacios et al., 2024). Aunque es posible reconstruir de forma general el tipo de tecnología empleada (Watson Explorer/Content Analytics) y su finalidad operativa (búsqueda y asociación en el SPOA), no se dispone públicamente de la lógica interna que explique por qué un caso se asocia con otro ni de métricas completas de desempeño, error o sesgo por subpoblaciones (Palacios et al., 2024).

En trazabilidad, los reportes de gestión documentan resultados agregados (por ejemplo, cantidad de casos asociados o productos analíticos generados), pero no hacen visible el sistema operativo necesario para auditar consultas, criterios aplicados y rutas de decisión internas en cada operación concreta. En consecuencia, la trazabilidad se caracteriza como parcial: documentable a nivel de existencia de reportes y productos institucionales, pero insuficiente para habilitar una auditoría externa completa del proceso de transformación de datos en asociaciones (Palacios et al., 2024).

6.5.10. Control humano significativo

El diseño de uso descrito para Fiscal Watson presupone intervención humana en la formulación de consultas, interpretación de resultados y verificación posterior, dado que el output se orienta a apoyar hipótesis investigativas y no a automatizar decisiones jurídicas finales. En la operación, el control humano se materializa en equipos de analistas que contrastan resultados de asociación, depuran registros y articulan insumos para dependencias

investigativas, lo que introduce un nivel de supervisión durante el uso (Fiscalía General de la Nación, 2020). Ex post, el control depende de la posibilidad institucional de revisar resultados, ajustar parámetros de consulta y corregir registros del SPOA que afecten la calidad de la analítica (Palacios et al., 2024, p. 14). Así, el CHS existe en la operación práctica, pero condicionado por asimetrías de información respecto de la lógica propietaria y por la disponibilidad de registros que permitan revisar cómo se llegó a una asociación específica (Palacios et al., 2024).

6.5.11. Evidencia de impactos: afectaciones, contenciones y mecanismos de contención

En términos de resultados reportados, el informe de empalme institucional registra que la herramienta permitió asociar aproximadamente 25.000 números únicos de noticia criminal (NUNC) y reporta alrededor de 500 casos considerados “exitosos” en el marco del uso analítico para intervención temprana, lo que constituye la principal evidencia cuantitativa pública de impacto (Fiscalía General de la Nación, 2020,). El mismo documento enmarca estos resultados en un universo de información masiva del SPOA (del orden de 16 millones de registros), lo que refuerza la lectura institucional de la herramienta como mecanismo de gestión de escala frente a la saturación informacional (Fiscalía General de la Nación).

Desde la perspectiva de afectaciones y riesgos, el diagnóstico crítico disponible enfatiza que la ausencia de transparencia suficiente sobre criterios de asociación, métricas de error y gobernanza del dato puede traducirse en riesgos para garantías como el debido proceso y la igualdad material, especialmente si las asociaciones influyen en priorización o atribución de relevancia investigativa (Palacios et al., 2024). Como mecanismos de

contención, las fuentes públicas permiten identificar principalmente controles organizacionales (intervención de analistas y prohibición implícita de automatización total) más que salvaguardas normativas específicas o evaluaciones de impacto integrales publicadas, lo que sugiere un esquema de mitigación limitado a la práctica interna y a reportes ex post (Fiscalía General de la Nación, 2020).

6.5.12. Síntesis del caso

Una vez realizada la reconstrucción técnico institucional del caso, a continuación se presenta a manera de resumen los hallazgos de la caracterización técnico-institucional de forma estructura como se había establecido previamente en la metodología Tabla 1 cuya función es condensar en un formato homogéneo los elementos relevantes del sistema analizado (finalidad, población impactada, datos, tratamiento, modelo, salida, rol, gobernanza, transparencia/trazabilidad, control humano significativo y evidencia de impactos), permitiendo visualizar de manera integrada cómo se articula el dispositivo y dónde se ubican los principales puntos de tensión.

Tabla 8
Síntesis de la caracterización técnico-institucional del caso Fiscal Watson

Dimensión	Preguntas guía	Hallazgos clave
Finalidad declarada y contexto	¿Qué problema pretende resolver? ¿En qué decisión/incidencia estatal se inserta?	Fiscal Watson es una herramienta de analítica de contenidos y búsqueda avanzada, basada en tecnología IBM Watson, implementada por la Fiscalía para apoyar la investigación penal mediante el hallazgo y asociación de información del SPOA. Su finalidad declarada es procesar grandes volúmenes de registros y

Dimensión	Preguntas guía	Hallazgos clave
		textos para identificar relaciones entre noticias criminales y producir insumos de gestión y priorización, sin sustituir formalmente el juicio del fiscal o investigador.
Población/territorio impactado	¿A quién aplica? ¿Se focaliza territorial o poblacionalmente?	El sistema impacta principalmente a quienes aparecen vinculados a noticias criminales del SPOA (víctimas, indiciados, imputados y terceros mencionados), pues sus datos pueden ser indexados y correlacionados para construir hipótesis de conexión. Territorialmente, opera sobre información de seccionales y dependencias en todo el país, con alcance potencialmente interjurisdiccional. La intensidad del impacto varía según la fase procesal, la dependencia usuaria y el tipo de consulta aplicada.
Datos: origen y calidad	¿De dónde vienen los datos? ¿Son pertinentes, representativos, actualizados?	Fiscal Watson se alimenta sobre todo del SPOA, con datos estructurados y no estructurados (denuncias, informes, actuaciones, relatos textuales y metadatos) susceptibles de indexación. Para asociar casos usa variables como georreferenciación, lugares y fechas, patrones o modus operandi e indicadores vinculados a reincidencia. La calidad depende de completitud y estandarización del registro, y algunas variables (territorio, historial, redes) pueden operar como proxies socioeconómicos o de selectividad policial.
Tratamiento de datos	¿Cómo se recolectan, cruzan, conservan y comparten?	El tratamiento parte de la consolidación en el SPOA y continúa con extracción e indexación para habilitar búsquedas complejas y cruces por coincidencias textuales, temporales y espaciales. Hay

Dimensión	Preguntas guía	Hallazgos clave
		reconocimiento judicial de su uso como soporte de búsqueda sobre bases del SPOA. Sin embargo, la información pública no permite precisar retención, conservación, accesos granulares, logging ni intercambio interagencial; por ello, estos flujos deben registrarse como no verificables públicamente.
Modelo/arquitectura	¿Qué tipo de modelo (reglas/ML/DL)? ¿Qué variables usa?	Fiscal Watson opera como sistema de recuperación y descubrimiento de información: indexa contenido, extrae entidades y habilita consultas avanzadas para correlacionar textos y metadatos, más que predecir conductas. Se identifican componentes IBM Watson Explorer/Content Analytics, pero la especificación de los algoritmos de asociación (ponderaciones, reglas, ajustes de relevancia) no es auditable con documentos públicos y permanece opaca por su carácter propietario.
Salida del sistema	¿Qué produce (score, alerta, recomendación)?	Su output se expresa principalmente en resultados de búsqueda y reportes de asociación de noticias criminales dentro del SPOA, permitiendo explorar vínculos por variables espaciales, temporales y de modus operandi. Institucionalmente se reporta un índice de noticias asociadas para orientar asignación de recursos. Inferencialmente, los resultados deben leerse como hipótesis de relación basadas en patrones del corpus, no como conclusiones probatorias autosuficientes; requieren verificación humana posterior.
Rol en la decisión	¿La salida decide, recomienda o prioriza?	Formalmente, Fiscal Watson funciona como insumo para análisis, priorización y asignación de recursos

Dimensión	Preguntas guía	Hallazgos clave
		investigativos, sin sustituir la decisión del fiscal. En la práctica, puede orientar qué casos se conectan y qué hipótesis se exploran. El riesgo es la automatización de facto: que la salida se naturalice como criterio predominante por eficiencia organizacional. La intensidad del efecto depende de la dependencia institucional y de los controles internos de uso.
Gobernanza, evaluación de impacto y responsabilidades	¿Quién opera? ¿Quién responde? ¿Qué controles y evaluaciones existen?	La gobernanza se compone de proveedor tecnológico (IBM) y entidad operadora (Fiscalía), que define finalidades, escenarios y responsables internos. Los reportes lo enmarcan en estrategias de big data e innovación. No obstante, la evidencia pública no permite verificar evaluaciones integrales ex ante en derechos. Así, las responsabilidades quedan fragmentadas entre diseño y actualización técnica, administración del dato y decisiones sobre el uso del output, con control más reactivo que preventivo.
Transparencia y trazabilidad	¿Es auditable? ¿Hay explicabilidad/registro?	La transparencia es limitada por la naturaleza privada del activo digital y escasa documentación pública sobre criterios de asociación y protocolos. Puede identificarse la tecnología y finalidad general, pero no es visible por qué se asocia un caso con otro ni hay métricas completas de desempeño, error o sesgo. En trazabilidad, los informes muestran resultados agregados, sin el logging operativo necesario para auditar consultas, criterios aplicados y rutas de decisión.
Control humano significativo (CHS)	¿Dónde interviene el humano (diseño-uso-revisión)?	El control humano significativo se expresa en la formulación de consultas, interpretación y

Dimensión	Preguntas guía	Hallazgos clave
		verificación: analistas contrastan asociaciones, depuran registros y elaboran insumos para equipos investigativos. Ex post, el control dependería de revisar resultados, ajustar parámetros y corregir registros del SPOA, aunque los mecanismos específicos no son plenamente verificables públicamente. En suma, el CHS existe, pero queda condicionado por la opacidad propietaria y la disponibilidad de trazas revisables.
Evidencia de impactos	¿Hay evidencia de sesgos/errores/afectaciones?	Como evidencia pública, se reporta la asociación de cerca de 25.000 NUNC y alrededor de 500 casos "exitosos", en un universo de información masiva del SPOA. En riesgos, diagnósticos críticos señalan que la falta de transparencia sobre criterios de asociación, métricas de error y gobernanza del dato puede afectar debido proceso e igualdad si influye en priorización o relevancia investigativa. Las contenciones visibles son principalmente organizacionales y ex post.

7. Capítulo III. Estudio comparado y umbrales de validez constitucional

El objetivo de este capítulo es reconstruir los umbrales mínimos de aceptabilidad constitucional que deben regir la incorporación de Inteligencia Artificial en funciones estatales de control y vigilancia, a partir de las tensiones identificadas en la caracterización técnico-institucional de cada sistema. No se trata todavía de juzgar la validez definitiva de cada caso, tarea reservada al Capítulo IV, sino de derivar, desde sus fallas de diseño,

gobernanza e impacto, los criterios jurídicos indispensables para que tales herramientas no erosionen las garantías del Estado Social de Derecho

7.1 Caso COMPAS: riesgo actuarial y desigualdad estructural

El análisis constitucional de COMPAS exige situar la herramienta más allá de su descripción técnica. Más que un recurso neutro de optimización, COMPAS puede leerse como un dispositivo que reconfigura las garantías del Estado Social de Derecho al desplazar la valoración singular del sujeto hacia perfiles estadísticos de riesgo.

7.1.1. Igualdad Material

El umbral de la igualdad material exige que el Estado no se limite a una paridad formal, sino que remueva activamente los obstáculos que producen desigualdades de hecho (Corte Constitucional, Sentencia C-371/00). En COMPAS, ese umbral se tensiona porque el sistema opera bajo una lógica de pasado como patrón, donde las desventajas históricas de comunidades racializadas se plasman en los datos y se refuerzan en el modelo como si fueran una situación objetiva, sin una justificación estricta y proporcional.

La evidencia técnico-institucional muestra que COMPAS actúa sobre poblaciones ya sometidas a hipervigilancia penal, y concentra sobre ellas cargas adicionales y asimétricas de vigilancia. Al ignorar que los datos administrativos son construcciones sociales y no hechos neutros, el sistema corre el riesgo de tratar igual a quienes parten de situaciones materialmente desiguales y de consolidar la marginalización bajo un ropaje de eficiencia técnica (Mayson, 2018).

La aparente neutralidad técnica del sistema no neutraliza ese problema. La disparidad documentada en falsos positivos (44.9% en población negra frente a 23.5% en población blanca) muestra que un sistema puede exhibir consistencia estadística y, aun así, distribuir de manera desigual los costos del error sobre grupos históricamente vulnerados (Angwin et al., 2016a; Dieterich et al., 2016).

Desde esta perspectiva, la incorporación de la IA en la justicia penal, como en el caso COMPAS, representa un riesgo de deshumanización de la jurisdicción (Asís, 2023), ya que el sistema prioriza la eficiencia administrativa sobre la deliberación ética y humana necesaria para salvaguardar la dignidad de las personas (Hernández, 2025). Este fenómeno se inscribe en lo que algunos autores denominan la *trampa del formalismo*, en la cual la equidad se reduce a métricas estadísticas abstractas que, bajo un ropaje de objetividad, ignoran el contexto social, histórico y político de los grupos afectados (Coddou et al., 2025). Al operar como una *caja negra*, el sistema impide que el procesado comprenda o rebata la lógica de su clasificación de riesgo, lo que vulnera el derecho fundamental al debido proceso y el principio de transparencia algorítmica (Ibañez, 2023).

Asimismo, la implementación de estos sistemas bajo la lógica del pasado como destino genera un bucle de retroalimentación donde los algoritmos, al entrenarse con datos de sentencias y detenciones históricamente sesgadas, actúan como un blanqueo de sesgo (Solar, 2020). Esto significa que la IA no solo refleja el racismo estructural, sino que lo automatiza y expande, transformando prejuicios sociales en reglas matemáticas que cierran el futuro del individuo a una mera extrapolación de las desventajas de su entorno (Innerarity. 2025). Por tanto, se produce una fractura entre la justicia formal, centrada en la precisión

estadística, y la justicia material, que exige una interpretación cualitativa y humana de los hechos y valores que el algoritmo es incapaz de procesar por sí mismo. (Gómez et al., 2025).

Para que el uso de la IA sea constitucionalmente legítimo, es imperativo que estas herramientas mantengan un “rostro humano", actuando siempre como un instrumento de apoyo auxiliar y nunca sustitutivo de la labor del juez (Gómez et al., 2025). Solo a través de una supervisión humana efectiva y del reconocimiento de la autonomía del decisor se puede evitar que el sujeto sea reducido a un simple paquete de datos y se garantice que la tecnología sirva a la justicia social en lugar de consolidar nuevas formas de dominación algocrática (Flórez, 2025).

7.1.2. No discriminación: El poder de los Proxies y la Vigilancia Biopolitica

Esta garantía demanda la identificación y mitigación de la discriminación indirecta, la cual ocurre cuando prácticas en apariencia neutras influyen de manera desproporcionada en grupos protegidos (ACNUDH, 2022). Dicho en otras palabras, la no discriminación prohíbe cualquier distinción basada en criterios sospechosos (raza, condición social, etc.) que anule el ejercicio de derechos (Comité de Derechos Económicos, Sociales y Culturales, Observación General Nº 20, 2009). En la era algorítmica, esta garantía enfrenta la discriminación indirecta producida por el uso de *proxies*.

Y es que el análisis crítico de COMPAS revela que, aunque el sistema es formalmente se declare ciego a la raza, utiliza múltiples proxies (historial de arrestos, empleo, estabilidad residencial, nivel educativo) actúan como espejos de las disparidades estructurales previas (Mayson, 2018). Desde la sociología del castigo, esto constituye una forma de vigilancia biopolítica en donde el poder se ejerce mediante la clasificación y segmentación de la

población según su riesgo y que consecuentemente normaliza la sospecha sobre ciertos cuerpos (Garland, 1999). La disparidad documentada en las tasas de error —donde los acusados negros reciben casi el doble de falsos positivos (44.9%) que los blancos (23.5%)— evidencia que el algoritmo proyecta el sesgo policial histórico hacia el futuro, convirtiendo la predicción en una profecía autocumplida que castiga la pertenencia a un grupo y no la conducta individual (Ibáñez, 2020).

7.1.3. Dignidad Humana: Contra la Algocracia

La dignidad humana funciona como un límite infranqueable que prohíbe la cosificación o instrumentalización de la persona (Atienza, 2022). COMPAS erosiona esta garantía al reducir la complejidad de la experiencia humana a un puntaje o score en deciles (1–10), convirtiendo al individuo —según el propio manual operativo de la empresa desarrolladora— en un objeto de gestión actuarial (Equivant, 2017).

En esta dinámica de poder, el sujeto deja de ser concebido como un fin en sí mismo para transformarse en un *excedente conductual* procesado por el capitalismo de la vigilancia con fines de control estatal (Zuboff, 2019, p. 237). La imposición de etiquetas de riesgo basadas en datos agregados de grupos similares y no en la certeza individual sustituye la valoración de la singularidad humana por una probabilidad estadística, lo que constituye una forma de despersonalización incompatible con el Estado Social de Derecho (Atienza, 2022).

Esta delegación de la toma de decisiones judiciales a sistemas automatizados, conocida como algocracia, altera los cimientos de la legitimidad legal al reemplazar la deliberación ética por la opacidad sistémica (Flórez, 2025). Al operar bajo lógicas que desvían la atención de la singularidad biográfica, el sistema transforma al ciudadano en un

paquete de datos consumible, despojándolo de su condición de ser social para reducirlo a un individualismo en red manejable por el poder estatal o corporativo (Barrios Tao, 2024). Esta cristalización de identidades mediante categorizaciones fijas y a menudo inmodificables debilita la libertad de autodeterminación, insertando al sujeto en un régimen de vigilancia del que resulta prácticamente imposible negociar su propia representación (Santana, 2025).

En consecuencia, la autonomía personal deja de ser un derecho protegido para convertirse en una variable ajustada a la conveniencia de terceros, lo que compromete gravemente el postulado del libre desarrollo de la personalidad al inducir a las personas a ser "sujetos pensados" en lugar de "sujetos pensantes" (Santana, 2025). La primacía de la eficacia y la celeridad administrativa no puede justificar el sacrificio de los derechos inalienables, pues la justicia debe ser, ante todo, una tecnología centrada en el ser humano y no un instrumento de deshumanización que reduzca a las personas a meros puntos de datos estadísticos (Galera, 2024). Finalmente, el hermetismo de la caja negra algorítmica entorpece la imputación de responsabilidades y la transparencia necesaria en un Estado de Derecho, obligando a los ciudadanos a depositar una fe ciega en sistemas cuya lógica interna es ininteligible incluso para los propios operadores jurídicos (Hernández Terán, 2024).

7.1.4. El control humano significativo

El control humano significativo (CHS) es un derecho instrumental indispensable para asegurar que las decisiones automatizadas permanezcan bajo la órbita de responsabilidad humana (Sánchez, 2021). No se satisface con una intervención decorativa, sino que exige la capacidad real del operador para cuestionar, modular o revertir el resultado.

En COMPAS, el CHS se ve severamente degradado por el fenómeno del sesgo de automatización, donde la celeridad administrativa y la apariencia de experticia del algoritmo llevan a la autoridad que toma decisiones a confiar ciegamente en el número (Washington, 2019). Además, la configuración del sistema como una caja negra protegida por secretos comerciales impide que el operador humano comprenda realmente la lógica de la ponderación interna, lo que reduce el control a una función meramente decorativa o formal. Contrariamente, para que el control sea significativo, el decisor debe tener la capacidad real de contradecir el output, lo cual es materialmente imposible si no existe explicabilidad y trazabilidad plena del proceso que transformó los datos en una etiqueta de riesgo (Washington, 2019; Sánchez, 2021).

En consecuencia, COMPAS revela una transición hacia la algocracia, donde el poder ya no se ejerce mediante normas generales y públicas, sino a través de arquitecturas de datos opacas que normalizan la vigilancia y reproducen, de forma silente, las asimetrías de poder que el Estado Social de Derecho debería corregir. Este dispositivo no solo automatiza la gestión del riesgo, sino que funciona como un espejo del pasado que reproduce, de forma silente, las asimetrías de poder que el Estado Social de Derecho tiene el mandato de corregir (Mayson, 2018).

Pese a ser un caso radicado en los Estados Unidos, su análisis es fundamental para Colombia porque actúa como una advertencia constitucional frente al despliegue de herramientas locales como PRISMA o el Fiscal Watson, los cuales comparten la misma lógica actuarial de traducir conflictos sociales en perfiles estadísticos. La experiencia de COMPAS demuestra que, en contextos de debilidad regulatoria y crisis penitenciaria como el colombiano, la introducción acrítica de estos modelos corre el riesgo de consolidar fines

encubiertos de la pena, tales como la estigmatización y exclusión de sectores vulnerables, todo lo cual lo lleva bajo la apariencia de una eficiencia administrativa neutra.

Por tanto, COMPAS provee el estándar de contraste necesario para exigir que cualquier IA estatal en Colombia no solo cumpla con una precisión técnica mínima, sino que se someta a un juicio de igualdad material (métricas de equidad) que impida que el algoritmo se convierta en un instrumento de dominación que erosione la dignidad humana y la autonomía cognitiva del sujeto procesado.

7.2 El Caso SyRI: Del Panóptico Disciplinario a la Gubernamentalidad Algorítmica o vigilancia social y sospecha automatizada

El despliegue del sistema System Risk Indication (SyRI) en los Países Bajos no constituyó un mero avance administrativo, sino que simboliza el tránsito hacia un nuevo régimen de poder. Esta transición marca el paso del panóptico tradicional, que Foucault describió como un mecanismo de vigilancia física y corrección del cuerpo (Foucault, 1977), hacia una gubernamentalidad algorítmica que opera en la preconsciencia del ciudadano. En este nuevo paradigma, el Estado ya no busca disciplinar mediante la visibilidad constante, sino gobernar mediante la invisibilidad del dato y la previsión estadística.

Bajo esta lógica, el poder estatal se ejerce mediante el procesamiento masivo de información para predecir y perfilar conductas, lo que Shoshana Zuboff denomina la captura del excedente conductual para fines de certidumbre total (Zuboff, 2020). El caso SyRI ilustra cómo la tecnología es instrumentalizada para anular la contingencia de lo humano, limitando la capacidad del individuo para actuar de forma imprevista o comenzar algo nuevo, cualidad que constituye la esencia de la libertad política (Arendt, 1998, citado en Innerarity, 2025). Al

basar sus decisiones en patrones del pasado, estos sistemas congelan al sujeto en sus huellas digitales, negándole el derecho a una identidad dinámica y evolutiva (Innerarity, 2025).

Uno de los aspectos más críticos de este sistema es su opacidad intrínseca o el fenómeno de la caja negra. La falta de transparencia sustantiva impide que el afectado comprenda la lógica del algoritmo, convirtiendo la decisión administrativa en un veredicto inquebrantable que hace inviable el ejercicio del derecho de defensa (Pérez Conchillo, 2025). Esta asimetría cognitiva no es accidental, sino que forma parte de una arquitectura de control donde la verdad es de naturaleza retroactiva y se impone sobre la realidad social (Innerarity, 2025).

Asimismo, la selectividad discriminatoria de SyRI evidencia la automatización de la desigualdad. El Tribunal de Distrito de La Haya declaró que la normativa que regulaba SyRI no tiene efecto vinculante por violar el derecho a la privacidad (Art. 8 del CEDH), al constatar que el sistema se aplicaba exclusivamente en "distritos problemáticos" y carecía de la transparencia necesaria para garantizar que no se discriminara a poblaciones vulnerables basándose en su estatus socioeconómico o su origen (Rechtbank Den Haag, 2020). Este fenómeno, catalogado por Ruha Benjamin (Benjamin, 2019, citado en Duque 2025) como el *New Jim Code*, demuestra que los algoritmos no son neutros, sino que reproducen y amplifican sesgos raciales y socioeconómicos incrustados en los datos históricos em donde la eficiencia tecnológica se convierte así en un pretexto para la exclusión social, bajo un manto de supuesta objetividad científica.

Ante este panorama, las garantías del Estado Social de Derecho exigen una profunda reinterpretación. No basta con regular el tratamiento de datos personales; es imperativo consagrar un nuevo paradigma de protección que reconozca el derecho a una decisión

humana significativa y la construcción de una identidad ajena a las decisiones algorítmicas. El ordenamiento jurídico debe evolucionar desde la mera "transparencia de código" hacia una "transparencia de diseño" que permita auditar los valores y fines políticos que guían al sistema (Innerarity, 2025). Finalmente, es necesaria la formalización del control humano sobre la inteligencia artificial como un nuevo Derecho Humano, garantizando que el ciudadano permanezca como sujeto titular de derechos y no sea reducido a un mero objeto de la mirada estadística (Hernández et al, 2025).

7.2.1. Igualdad material: La asimetría en la carga de la vigilancia estatal de SyRI

La igualdad material es un mandato de optimización que exige al Estado no solo abstenerse de tratos discriminatorios, sino remover activamente los obstáculos socioeconómicos que impiden el goce de derechos de los grupos marginados (Corte Constitucional, 2020). No obstante, en la era de la gubernamentalidad algorítmica, este principio se ve comprometido por diseños tecnológicos que, bajo una supuesta neutralidad, concentran de forma opaca las cargas de sospecha en poblaciones históricamente vulneradas (Santana Ramos, 2025). Estos sistemas no se limitan a procesar información, sino que instituyen una nueva arquitectura de control que redefine quién es merecedor de confianza y quién es catalogado como un riesgo administrativo.

Esta focalización territorial implica que el Estado asume la pobreza como un indicador inherente de riesgo, rompiendo el equilibrio relacional de la igualdad (Sentencia C-084/20) y sustituyendo la presunción de inocencia por una "presunción de fraude" basada en la pertenencia a un código postal determinado (Innerarity, 2025). Al operar de esta manera,

el algoritmo no solo descubre la realidad, sino que la prescribe, congelando a los sujetos en perfiles estadísticos de los que no pueden escapar (Innerarity, 2025).

Esta dinámica consolida lo que se puede denominar como vigilancia de clase, puesto que al operar exclusivamente sobre quienes dependen de la red de seguridad social, el sistema impone una carga de transparencia asimétrica; mientras los ciudadanos con recursos mantienen su privacidad, los más pobres son "sospechosos de antemano" por el solo hecho de solicitar asistencia (Alston, 2019). Esta dinámica consolida una desigualdad de hecho donde la precariedad económica activa dispositivos de control que otros sectores eluden (Corte Constitucional, 2000).

En consecuencia, la tecnología se convierte en un instrumento de biodeterminación que impone una disciplina punitiva sobre los más débiles, permitiendo que los sectores opulentos eludan los dispositivos de control que la administración automatizada activa exclusivamente para la gestión de la pobreza.

Finalmente, esta automatización de la desigualdad evidencia que los algoritmos como SyRI, lejos de ser objetivos, actúan como mecanismos de blanqueo del sesgo, transformando prejuicios históricos en veredictos tecnológicos supuestamente científicos (Innerarity, 2025). Para que el Estado Social recupere su esencia garante, es imperativo que los sistemas de decisión automatizada sean sometidos a un escrutinio que no solo audite el código, sino que evalúe su impacto dispar y su capacidad para perpetuar estructuras de opresión de clase (Coddou McManus et al., 2025)

7.2.2. No discriminación

El derecho a la no discriminación constituye un pilar fundamental que prohíbe cualquier distinción basada en criterios sospechosos que anule el ejercicio de derechos fundamentales (ACNUDH, 2022, 1). No obstante, la discriminación algorítmica se manifiesta de forma particularmente perniciosa al presentarse como una modalidad indirecta o sistémica, camuflada bajo una supuesta neutralidad técnica que, en realidad, actúa como un blanqueo del sesgo (Innerarity, 2024). Este fenómeno no solo reproduce las injusticias del mundo analógico, sino que las dota de una apariencia de verdad incontestable, dificultando el escrutinio público y el ejercicio del derecho de defensa (Pérez Conchillo, 2025).

Un elemento crítico en esta arquitectura es el peligro de las variables sustitutas o *proxies*. Aunque sistemas como SyRI eviten el uso de categorías explícitamente prohibidas, como la raza o el origen étnico, procesan redes de datos aparentemente neutros tales como, vivienda, deudas, educación, que funcionan como indicadores indirectos de la condición socioeconómica y el origen étnico (Lazcoz & Castillo, 2020). Esta correlación permite al algoritmo identificar y penalizar a colectivos específicos sin nombrarlos, perpetuando estereotipos discriminatorios bajo un manto de racionalidad científica (, Galera Victoria, 2025). Esta "biodeterminación de la credibilidad" desplaza la presunción de inocencia por una sospecha automatizada que se ensaña con quienes habitan en los márgenes del sistema (Romano, 2025).

Esta dinámica se agrava mediante lo que se denomina el bucle de exclusión o sesgo de retroalimentación. La utilización de datos históricos sesgados para entrenar modelos predictivos genera un ciclo de retroalimentación donde se vigila más a quienes ya han sido controlados, convirtiendo el prejuicio en una profecía autocumplida (Lazcoz & Castillo, 2020). Así, el sistema no detecta el fraude de manera neutral y objetiva, sino que automatiza

la exclusión estructural que envía a las fuerzas de control a vecindarios con presencia excesiva de vigilancia previa, independientemente de la tasa real de criminalidad (Ibáñez López-Pozas, 2025).

Bajo este paradigma, el futuro pierde su carácter abierto para convertirse en una extrapolación ininterrumpida del pasado, donde el sujeto queda congelado en el "eterno ahora" de sus desventajas estructurales (Innerarity, 2024, 1217). Para que el Estado Social de Derecho pueda enfrentar este New Jim Code, es imperativo trascender la transparencia formal del código y exigir una transparencia de diseño que audite no solo los datos, sino los valores políticos y las jerarquías que el sistema pretende naturalizar. La discriminación algorítmica, por tanto, no debe verse como un error técnico corregible, sino como un objeto sociotécnico que refleja y amplifica las relaciones de poder de la sociedad que lo construye (Ramírez, 2023).

7.2.3. Dignidad Humana

La dignidad humana es el fundamento del orden político y exige que el individuo sea tratado siempre como un fin en sí mismo y nunca exclusivamente como un medio para la eficiencia del Estado (Atienza, 2022). Bajo este imperativo, el Estado Social de Derecho debe garantizar que la persona pueda desplegar su proyecto de vida con autonomía y sin verse sometida a humillaciones que anulen su subjetividad moral (Santana Ramos, 2025; Sentencia T-033/24). Sin embargo, el despliegue de sistemas de vigilancia como SyRI pone en entredicho este principio al instrumentalizar la existencia del ciudadano en favor de una lógica de "certidumbre total" que sacrifica la libertad en aras de la previsión estadística (Zuboff, 2020, 354).

En este contexto, SyRI incurre en una cosificación y estigmatización sistémica al reducir la complejidad de la vida humana a una simple "notificación de riesgo" o puntuación algorítmica. Este proceso de "datificación" transforma al ser humano de un sujeto titular de derechos en un objeto de gestión punitiva (Lazcoz & Castillo, 2020)., donde su identidad es fragmentada y reconstruida por máquinas que operan bajo parámetros inescrutables (Santana Ramos, 2025). Al delegar el juicio sobre la honestidad de un ciudadano a un sistema automatizado, el Estado renuncia a reconocer la singularidad de la persona, reemplazando al "sujeto pensante" de la modernidad por un "sujeto pensado" y predeterminado por el rastro de sus datos (Quiceno, 2025).

Asimismo, la opacidad intrínseca de estos sistemas opera como una forma de humillación administrativa que anula la autodeterminación informativa (Lazcoz & Castillo, 2020). La negativa del Estado a revelar los indicadores de riesgo y la lógica del algoritmo sumerge al ciudadano en un estado de indefensión absoluta frente a la caja negra estatal, impidiéndole comprender y cuestionar las premisas que fundamentan la sospecha en su contra.

Como sostiene Atienza (2022), la dignidad implica el derecho a conocer y cuestionar aquello que nos afecta; de lo contrario, se anula la subjetividad moral y el individuo queda en un estado de indefensión absoluta frente a la "caja negra" estatal. Esta asimetría cognitiva genera una dependencia epistémica que erosiona la confianza en las instituciones y aliena al individuo de los procesos de deliberación que afectan su vida.

Por tanto, la dignidad en la era digital no debe ser entendida solo como el respeto a la privacidad, sino como el derecho a no ser reducido a un perfil estadístico inalterable. La transparencia de diseño se vuelve entonces un requisito de justicia material, permitiendo que

la persona recupere su capacidad de interlocución frente al poder (Innerarity, 2024). Sin mecanismos de control humano significativo que garanticen la explicabilidad de las decisiones, la gobernanza algorítmica corre el riesgo de convertir la administración de lo social en una colmen" donde la autonomía personal es sacrificada en el altar de una eficiencia deshumanizada (Zuboff, 2020). El respeto a la dignidad humana exige, en última instancia, que el futuro del individuo permanezca como un espacio de contingencia y libertad, y no como una sentencia dictada por los datos del pasado (Innerarity, 2024, 1229).

7.2.4. Control Humano Significativo

El control humano significativo (CHS) no debe entenderse como un requisito técnico periférico, sino como una condición instrumental indispensable que asegura que las decisiones algorítmicas permanezcan bajo la protección de la responsabilidad institucional y la ética pública. En un Estado Social de Derecho, la delegación de decisiones a sistemas automatizados sin un CHS real convierte las garantías de igualdad y dignidad en enunciados incontrolables, pues el sistema opera bajo una lógica de caja negra que elude el escrutinio humano (Gómez Pavajeau, 2024). Este control exige que la persona encargada de la supervisión sea un agente moral capaz de comprender, validar y, de ser necesario, anular la recomendación de la máquina para preservar los valores constitucionales (Innerarity, 2024).

Uno de los hallazgos más críticos del caso SyRI fue la emergencia de lo que se denomina el humano decorativo. En la implementación del sistema neerlandés, la participación de los funcionarios fue meramente formal, ya que estos carecían de la competencia técnica y la información necesaria para comprender o disputar la lógica del modelo predictivo (Lazcoz & Castillo, 2020). Cuando el operador humano no comprende el

porqué de un resultado, la decisión no es humana, sino una delegación encubierta en la máquina (Van Bekkum & Zuiderveen Borgesius, 2021).

Esta figura es sintomática del sesgo de la automatización, donde el operador adopta una postura pasiva de confianza ilimitada en la máquina, perdiendo su capacidad crítica y actuando como un mero ejecutor de veredictos ininteligibles (Gómez Pavajeau, 2024).

En consecuencia, cuando el operador humano no comprende las razones de un resultado, no existe una decisión humana *stricto sensu*, sino una delegación encubierta y absoluta en la potencia de cálculo del algoritmo. Este vaciamiento de la función humana provocó el quiebre del "Fair Balance" o equilibrio justo. El Tribunal de Distrito de La Haya dictaminó que SyRI era ilegal al constatar que el sistema no lograba ponderar adecuadamente el interés público frente al derecho a la privacidad, debido precisamente a la ausencia de salvaguardas verificables que permitieran un control humano efectivo sobre el proceso de perfilado (Van Bekkum & Zuiderveen Borgesius, 2021). La ausencia de mecanismos de transparencia y explicabilidad impide que se identifiquen las variables de ponderación utilizadas, convirtiendo al algoritmo en una herramienta de dominación tecnocrática que se sitúa por encima de la rendición de cuentas administrativa (Lazcoz & Castillo, 2020).

Para restaurar la legitimidad democrática en el uso de la inteligencia artificial, el CHS debe evolucionar hacia lo que se denomina control humano por diseño**,** garantizando que el supervisor no sea un mero observador del ciclo final, sino un participante activo capaz de introducir la fricción necesaria frente a la eficiencia deshumanizada (Innerarity, 2024). De lo contrario, nos enfrentamos al riesgo de crear una zona moral deformada, donde los errores del sistema son atribuidos a funcionarios que, en la práctica, nunca tuvieron la capacidad real de influir en el resultado (Innerarity, 2024). El CHS, por tanto, se erige como la última

frontera del derecho para evitar que el ciudadano sea reducido a un objeto de gestión algorítmica sin posibilidad de contradicción o defensa (Hernández García de Velazco, 2025).

7.3 Caso AFR Locate: vigilancia biométrica y control policial

El análisis teórico-jurídico del sistema AFR Locate, operado por la Policía de Gales del Sur (South Wales Police - SWP), exige un examen riguroso de las tensiones entre la innovación tecnológica y las garantías fundamentales. El caso paradigmático *R (Bridges) v Chief Constable of South Wales Police* marca un hito en la fiscalización de la vigilancia biométrica y con ello permite examinar la tensión entre la eficiencia tecnológica y las garantías fundamentales del Estado Social de Derecho. A continuación, se desarrollan las garantías de dignidad humana, igualdad material, no discriminación y control humano significativo aplicadas a este caso.

7.3.1. Igualdad material

En el despliegue del sistema AFR Locate, la igualdad material se ve gravemente comprometida cuando el uso de tecnologías de vigilancia se proyecta de forma desproporcionada sobre colectivos específicos, transformando a los ciudadanos en tarjetas de identidad andantes. Desde una perspectiva postpositivista, el Derecho es una actividad dirigida al logro de fines y valores, donde la igualdad material no es solo un límite, sino un objetivo de justicia social innegociable (Atienza, 2022). En el caso *Bridges*, el tribunal puso de manifiesto como marcos legales con una discrecionalidad excesiva vulneran esta garantía, al carecer de criterios objetivos sobre el quién y el dónde se despliega la vigilancia, socavando la protección frente a la arbitrariedad estatal.

Y en el mismo sentido, uno de los hallazgos más relevantes realizado por la Corte de Apelación dentro de este litigio fue el de determinar la policía no cumplió con su deber de igualdad del sector público al no verificar de forma independiente si el software presentaba sesgos de raza o sexo. La evidencia técnica indica que los algoritmos de reconocimiento facial a menudo muestran tasas de falsos positivos desproporcionadamente altas para mujeres y minorías étnicas.

La tecnología, lejos de ser neutra, tiende a reproducir lógicas discriminatorias sistémicas arraigadas en la sociedad. En AFR Locate, la selección de los lugares de despliegue y la composición de las listas de vigilancia (*watchlists*) pueden actuar como *proxies* de condiciones socioeconómicas, concentrando la vigilancia en zonas o poblaciones específicas y reforzando trayectorias de exclusión. Y es que al concentrar la vigilancia en zonas de alta vulnerabilidad, el sistema no solo detecta el riesgo, sino que lo predetermina, reforzando trayectorias de exclusión que marginalizan aún más a las poblaciones históricamente castigadas (Ibáñez 2025). Como advierte Daniel Innerarity, cuando los algoritmos adoptan el pasado desigual como patrón inalterable para el futuro, el sujeto queda atrapado en un "eterno ahora" de estigmatización algorítmica (Innerarity, 2025).

Desde una óptica sociológica, estos sistemas instituyen una identificación singularizante que despoja al individuo de su anonimato en el espacio público, condicionando su libertad de movimiento y su derecho a la autodeterminación. La igualdad material, por tanto, se ve anulada si el diseño tecnológico no asume un enfoque proactivo para eliminar la discriminación sistémica arraigada en los datos históricos (Innerarity, 2025). Para que el Estado Social recupere su función garante, es imperativo que estos sistemas sean sometidos a evaluaciones de impacto de equidad que trasciendan la eficiencia operativa y

garanticen que la tecnología permanezca como un instrumento al servicio de la dignidad humana y no como una herramienta de dominación tecnocrática (Hernández García de Velazco, 2025).

7.3.2. No discriminación

En el caso de AFR Locate, la Corte de Apelación determinó que la policía de Gales del Sur no cumplió con su Deber de Igualdad del Sector Público al no investigar si el software presentaba sesgos inherentes (Bridges v. South Wales Police, 2020).

La evidencia técnica indica que los algoritmos de reconocimiento facial a menudo muestran tasas de falsos positivos desproporcionadamente altas para mujeres y minorías étnicas (Big Brother Watch, 2018). Esta discriminación indirecta ocurre cuando leyes o prácticas en apariencia neutras influyen de manera desproporcionada en los derechos de las personas, al entrenarse con datos que reflejan un pasado desigual, de manera que el sistema "aprende" a reproducir estigmas bajo un ropaje de objetividad algorítmica (Asís, 2023). En AFR Locate, la opacidad del algoritmo impidió que los ciudadanos afectados pudieran verificar o impugnar la imparcialidad del sistema, consolidando una forma de dominación técnica (Davies et al., 2018).

7.3.3. Dignidad Humana

En el caso de AFR Locate, la captura masiva y el procesamiento de datos biométricos en espacios públicos sin un consentimiento explícito plantea un riesgo de instrumentalización. Así, el uso de estas tecnologías puede convertir a los ciudadanos en tarjetas de identidad caminantes, donde el cuerpo físico es reducido a una estructura de

información procesable. Esta dinámica se inserta en lo que se denomina vigilancia biométrica masiva, la cual puede ser seductora e invisible, llevando a la ciudadanía a convertirse en una masa dócil gestionada por algoritmos.

Esta tecnología actualiza el panóptico foucaultiano, donde el poder ya no solo reprime, sino que produce sujetos dóciles mediante el registro y la normalización constante (Foucault, 2002). Al convertir a los ciudadanos en tarjetas de identidad caminantes, el sistema erosiona la esfera mental y el derecho al libre albedrío (Big Brother Watch, 2020). Atienza (2022) sostiene que la dignidad opera como un límite absoluto que prohíbe la degradación de la persona a la categoría de objeto. En este sentido, la vigilancia biométrica masiva de AFR Locate, al actuar de forma seductora e invisible, lleva a la ciudadanía a convertirse en una masa gestionada por algoritmos, afectando su autonomía cognitiva (Zuboff, 2019).

7.3.4. Control Humano Significativo

En la arquitectura técnica de AFR Locate, el control humano significativo (CHS) se implementa formalmente mediante un esquema de humano en el bucle (*human-in-the-loop*), donde toda alerta de coincidencia algorítmica debe ser validada por un operador antes de ejecutar una intervención policial. Sin embargo, este diseño se enfrenta al riesgo crítico de una automatización de facto, alimentada por lo que la doctrina denomina el sesgo de la automatización: una tendencia sistemática de los operadores a adoptar una postura pasiva de confianza ilimitada en las virtudes de la máquina, perdiendo su capacidad crítica de análisis (Gómez Pavajeau, 2024). Bajo presión operativa, el agente humano puede verse reducido a

un mero vehículo ejecutor de un veredicto previo del sistema, transformando la supervisión en un trámite puramente simbólico que anula la deliberación humana.

Este fenómeno se agrava por el denominado enigma del control (Innerarity, 2024)., que postula que cuanto más fiable parece un sistema automatizado, más difícil resulta para el supervisor mantener un nivel de compromiso cognitivo suficiente para intervenir eficazmente en caso de error. En el contexto de AFR Locate, la alerta algorítmica funciona como un anclaje que predetermina la decisión del oficial, creando lo que se define como una zona moral deformada (Innerarity, 2024), un escenario donde la responsabilidad jurídica de un posible error es atribuida al actor humano, a pesar de que el diseño del sistema y la opacidad técnica le impidieron influir de forma real en el resultado. De este modo, la presencia del humano no garantiza necesariamente la justicia del acto, sino que a menudo sirve para blindar la decisión de la máquina bajo un manto de aparente legalidad humana (Gómez Pavajeau, 2024).

Para que el control sea material y no meramente formal, el sistema debe ser transparente y explicable. El problema de la caja negra en AFR Locate reside en que, aunque el operador visualiza el *match*, carece de acceso a la lógica técnica fina de un algoritmo propietario protegido por derechos de propiedad intelectual, lo que le impide comprender las razones del acierto o la desviación estadística (Davies et al., 2018). Esta opacidad entra en conflicto con la necesidad de equilibrar la explotación comercial de la IA con la obligación de motivar y justificar las decisiones que afectan significativamente la vida de las personas (Cancio, 2025, 359). Sin una transparencia de diseño que permita auditar los parámetros de ponderación, el algoritmo se erige como una autoridad epistémica inescrutable que desplaza la racionalidad jurídica por una certeza computacional.

Desde una perspectiva iusfundamental, el CHS debe ser reinterpretado como un derecho instrumental indispensable en el Estado Social de Derecho, que asegura que la responsabilidad moral y jurídica permanezca en agentes humanos identificables y competentes (Gómez Pavajeau, 2024). Así diversos autores proponen la formalización del Control Humano de la Inteligencia Artificial como un nuevo Derecho Humano, derivado de la necesidad de corregir cualquier amenaza a la dignidad que surja de la delegación de funciones soberanas de vigilancia en sistemas opacos. En última instancia, el respeto a la autonomía humana exige que la tecnología permanezca como un recurso instrumental de apoyo y no como un sistema de toma de decisiones sustitutivo que despoje al ciudadano de su condición de sujeto titular de derechos (Innerarity, 2024, 1030).

7.4 Caso PRiSMA: justicia Predictiva

El análisis del sistema PRiSMA (Perfil de Riesgo de Reincidencia para la Solicitud de Medidas de Aseguramiento) exige un abordaje desde el Estado Social de Derecho, donde la tecnología no puede operar como un ente autónomo, sino como un dispositivo sujeto a los límites axiológicos de la Constitución. Implementado por la Fiscalía General de la Nación, este sistema funciona mediante un modelo de aprendizaje supervisado de Machine Learning que predice la probabilidad de reincidencia basándose en antecedentes penales y características del evento criminal. No obstante, este tránsito hacia una justicia actuarial desplaza el foco del derecho penal desde la culpabilidad por el acto cometido hacia la peligrosidad proyectada en el futuro, lo que tensiona el núcleo esencial de la presunción de inocencia. La "verdad" en este modelo deja de ser una reconstrucción del pasado para convertirse en una estimación estadística de la conducta venidera.

Uno de los impulsores de PRiSMA es la necesidad de racionalizar el uso de los cupos carcelarios, buscando que las medidas intramurales se concentren exclusivamente en sujetos con altos niveles de riesgo objetivo. Si bien este fin parece legítimo bajo una lógica de eficiencia administrativa, conlleva el riesgo de instrumentalizar al individuo, reduciendo su dignidad a una simple puntuación de riesgo en beneficio de la gestión del sistema penitenciario. Esta dinámica de cosificación algorítmica transforma al sujeto de derechos en un objeto de administración del peligro, donde su libertad depende de una arquitectura de datos que opera de forma preventiva. El peligro reside en que el Estado asuma que lo posible es idéntico a lo real, sustituyendo el juicio moral humano por una certidumbre computacional.

El desarrollo de PRiSMA afirmó haber excluido variables sociodemográficas para evitar sesgos por raza u origen étnico, centrándose únicamente en registros judiciales y penitenciarios. Sin embargo, la doctrina advierte que los datos históricos no son neutros y suelen reflejar prejuicios estructurales y disparidades socioeconómicas inherentes al sistema penal tradicional. Al entrenar el modelo con millones de registros de bases de datos de la Policía y la Fiscalía, existe el riesgo de crear un bucle de retroalimentación donde el algoritmo simplemente sistematiza y valida las desigualdades del pasado bajo un ropaje de objetividad técnica (Innerarity, 2024,). Por tanto, la exclusión formal de datos sensibles no garantiza per se la ausencia de una discriminación indirecta.

Finalmente, la legitimidad constitucional de PRiSMA depende estrictamente de que se preserve la no sustitución de la racionalidad humana. Como herramienta complementaria, el algoritmo debe servir de apoyo al fiscal y al juez, quienes nunca deben delegar su capacidad de discernimiento y valoración crítica en la máquina. El fenómeno del sesgo de la

automatización sugiere que el operador jurídico tiende a confiar ciegamente en la recomendación algorítmica, convirtiendo la intervención humana en un trámite simbólico. Para evitar una algocracia que erosione la confianza pública, es imperativo que los resultados de PRiSMA sean auditables, explicables y susceptibles de contradicción, garantizando que el futuro del imputado siga siendo un espacio de libertad y no una sentencia dictada por una caja negra estatal.

7.4.1. Igualdad materia en PRiSMA: El Sesgo de los Datos y la Discriminación Sistémica

La igualdad material, en el marco del Estado Social de Derecho, trasciende la mera paridad formal ante la ley para erigirse como un mandato de optimización que exige al Estado la remoción activa de los obstáculos que impiden a los grupos marginados el goce efectivo de sus derechos (Gómez Pavajeau, 2024). En el contexto del sistema PRiSMA, esta garantía se ve severamente tensionada por la naturaleza de sus datos de entrenamiento, los cuales provienen de registros de la Policía, la Fiscalía y el INPEC acumulados desde el año 2005 (Fiscalía General de la Nación, 2019,). Si estas bases de datos contienen las huellas de una vigilancia selectiva o prácticas policiales históricamente sesgadas contra sectores vulnerables, el algoritmo no predice el riesgo de forma neutra, sino que actúa como un espejo que reproduce y valida una discriminación sistémica bajo una apariencia de objetividad científica (Innerarity, 2024).

Un punto de conflicto central radica en la defensa institucional de la herramienta. La Fiscalía sostiene que PRiSMA garantiza la imparcialidad al aplicar los mismos criterios de evaluación a todos los sujetos y al haber excluido explícitamente variables

sociodemográficas como la raza o el origen étnico (López et al., 2023). No obstante, la igualdad material exige considerar que las condiciones socioeconómicas influyen determinantemente en la trayectoria penal de un individuo, y que ignorar el contexto de marginalidad al asignar un puntaje de riesgo puede convertir a la justicia en un dispositivo de gubernamentalidad algorítmica que castiga la exclusión previa.

Y es que, al basarse exclusivamente en la facticidad del pasado, el sistema incurre en una fuerza normativa de lo fáctico que congela al sujeto en sus desventajas estructurales, negándole la posibilidad de un futuro diferente a su registro estadístico (Innerarity, 2024). Esta dinámica se manifiesta con claridad en el fenómeno de los puntos calientes "hotspots" de criminalidad. Un individuo residente en un barrio periférico hipervigilado acumulará, por mera exposición, un mayor número de registros de capturas —incluso si estas no derivan en condenas, en comparación con un ciudadano de un estrato alto (Ibáñez, 2025). Al procesar esta información, sistemas como PRiSMA corren el riesgo de generar lo que Coddou Mc Manus et al. (2025) denominan un sesgo de retroalimentación: las clasificaciones de riesgo refuerzan la vigilancia sobre grupos ya estigmatizados, es decir, se vigila y captura más a quienes ya están en el sistema, convirtiendo el sesgo histórico en lo que otros autores llaman una profecía algorítmica autocumplida (Innerarity, 2024).

En consecuencia, el tratamiento de casos aparentemente similares bajo una misma métrica algorítmica puede derivar en un impacto dispar que profundiza la desigualdad social. La justicia predictiva, cuando se desprende de una hermenéutica consciente de las asimetrías de poder, corre el riesgo de sustituir la culpabilidad por el acto cometido por una peligrosidad estadística que recae desproporcionadamente sobre los más pobres (Innerarity, 2024).

Para que PRiSMA sea compatible con la igualdad material, es imperativo que su diseño no se limite a la eficiencia técnica, sino que incorpore mecanismos de transparencia sustantiva que permitan auditar cómo los datos del pasado están preconfigurando injustamente el destino de las poblaciones históricamente excluidas (Pérez Conchillo, 2025).

7.4.2. Garantía de No Discriminación: Del Sesgo Algorítmico a la Biodeterminación de la Credibilidad

El derecho a la no discriminación constituye una prohibición categórica contra cualquier distinción, exclusión o restricción que tenga por objeto anular o menoscabar el ejercicio de los derechos básicos de la persona (Santana Ramos, 2025, 9). En la era de la gubernamentalidad algorítmica, el mayor desafío para esta garantía no es la discriminación directa, sino la discriminación indirecta o sistémica, que ocurre cuando una práctica aparentemente neutra genera un impacto negativo desproporcionado en grupos históricamente protegidos (Coddou Mc Manus et al., 2025). Bajo esta lógica, la aparente objetividad del sistema PRiSMA debe ser cuestionada, pues la neutralidad matemática del algoritmo no garantiza la justicia de su resultado, pudiendo actuar más bien como un maquillaje tecnológico de prejuicios estructurales.

Un elemento de especial criticidad en PRiSMA es el peligro de las variables sustitutas o proxies. Aunque la Fiscalía General de la Nación sostiene que el sistema es "ciego" a categorías sensibles como la raza o el origen étnico (López et al., 2023), la herramienta procesa redes de datos sobre antecedentes judiciales, capturas previas y territorialidad que actúan como indicadores indirectos de la condición socioeconómica. En el contexto colombiano, estos factores suelen estar estrechamente correlacionados con la pertenencia a

minorías étnicas o estratos bajos, lo que permite al algoritmo identificar y penalizar a colectivos específicos bajo un manto de objetividad técnica. Esta dinámica reproduce el fenómenos de exclusión ya no se basadas en criterios raciales explícitos, sino en una criminalidad predeterminada por la vigilancia selectiva del pasado.

Asimismo, el sistema incurre en lo que puede denominarse la biodeterminación de la credibilidad (Romano, 2025)., al clasificar al individuo en un rango numérico de 0 a 100 basado en su probabilidad estadística de reincidencia (López et al., 2023). Esta fragmentación de la identidad reduce la complejidad biográfica del ser humano a un simple porcentaje o puntuación ciudadana, lo que conlleva una degradación del sujeto a objeto de gestión administrativa (Presno Linera, 2025).

Frente a esta algocracia que amenaza con automatizar la desigualdad, la jurisprudencia y los organismos internacionales, como el Comité de Derechos Económicos, Sociales y Culturales, exigen que los Estados adopten enfoques proactivos para eliminar la discriminación sistémica arraigada en las estructuras sociales (Galera, 2025). Para que PRiSMA sea legítimo en un Estado Social de Derecho, no basta con una transparencia formal del código; se requiere una transparencia de diseño que permita auditar cómo los datos del pasado están preconfigurando el futuro de los más vulnerables, garantizando que el juicio sobre la libertad humana permanezca siempre como una actividad deliberativa y no como un veredicto inescrutable de la máquina (Hernández García de Velazco, 2025).

7.4.3. Dignidad Humana y Justicia Predictiva: Entre la Autonomía y la Gestión Actuarial

La dignidad humana erige como el pilar fundamental que prohíbe la cosificación del individuo, imponiendo el imperativo ético de que toda persona sea tratada siempre como un fin en sí misma y nunca meramente como un medio para alcanzar la eficiencia del Estado (Atienza, 2022). Bajo esta premisa, el Estado Social de Derecho debe garantizar que el sujeto viva sin humillaciones y con la autonomía necesaria para definir su propio proyecto de vida, evitando que su existencia sea reducida a un cálculo de utilidad administrativa (Santana, 2025). Sin embargo, el despliegue de sistemas como PRiSMA pone en riesgo estos valores al instaurar una lógica de reificación y gestión actuarial, donde la complejidad de la trayectoria vital del procesado se comprime en una simple notificación de riesgo o puntuación algorítmica (Quiceno, 2025).

Al delegar la valoración de la peligrosidad a un cálculo probabilístico de reincidencia, el Estado renuncia a reconocer la singularidad del ser humano, tratándolo como un depósito de excedente conductual (Zuboff, 2020) cuya libertad es modulada para la optimización de los cupos carcelarios. Esta dinámica transforma al ciudadano, de un sujeto titular de derechos, a un objeto de administración del peligro, donde la decisión sobre su libertad no emana de un juicio moral individualizado, sino de una arquitectura de datos orientada a la certidumbre total (Gómez Pavajeau, 2024). En este escenario, la tecnología no actúa como un apoyo neutral, sino como un dispositivo que puede despojar a la persona de su identidad dinámica, congelándola en un perfil estadístico inalterable (Santana Ramos, 2025, 12).

Asimismo, el uso de PRiSMA genera una tensión insoluble entre el derecho penal de acto y el derecho penal de autor. Mientras que la Constitución colombiana opta por castigar lo que se hace y no lo que se es (Corte Constitucional sentencia C-226/02), la justicia predictiva se desliza hacia un modelo donde la anomalía social o la propensión detectada por

el algoritmo justifica preventivamente la privación de la libertad (Fair Trials, 2024). Este patrulleo algorítmico desplaza el foco de la culpabilidad por el hecho cometido hacia la sospecha sobre conductas futuras, lo cual resulta ontológicamente incompatible con el desarrollo de una autoidentidad. Al basar la intervención penal en perfiles de riesgo, el sistema impone una biodeterminación de la peligrosidad (Romano, 2025) que castiga al individuo por su pertenencia a un patrón estadístico y no por su responsabilidad individual.

Finalmente, el respeto a la autonomía y el libre desarrollo de la personalidad exige que el futuro del individuo permanezca siempre como un espacio de contingencia y libertad. La dignidad humana se vulnera entonces cuando el porvenir de una persona deja de ser una posibilidad abierta para convertirse en una sentencia dictada por los datos de su pasado y es que si el sistema penal permite que los algoritmos prefiguren el destino de los procesados, se anula la capacidad del sujeto para comenzar algo nuevo, cualidad que constituye la esencia misma de la condición humana (Innerarity, 2024, 121). En última instancia, una justicia que prioriza la predicción sobre la deliberación corre el riesgo de transformar la sociedad en una colmena donde la libertad es sacrificada en el altar de una eficiencia deshumanizada (Zuboff, 2020).

7.4.4. El Control Humano Significativo en PRiSMA: Responsabilidad Institucional frente al Sesgo de Automatización

El control humano significativo (CHS) se erige como una garantía instrumental inalienable dentro del Estado Social de Derecho, cuya función primordial es asegurar que la responsabilidad moral y jurídica de las decisiones públicas radique exclusivamente en

agentes humanos competentes y no en arquitecturas computacionales inescrutables. En el contexto de PRiSMA, esta garantía exige que el sistema no opere como un ente decisorio autónomo, sino como un recurso de apoyo que debe ser tamizado por la conciencia valorativa del funcionario judicial, evitando que la delegación tecnológica se traduzca en una abdicación de la soberanía jurídica.

No obstante, la implementación práctica de estos sistemas enfrenta el riesgo sistémico del sesgo de automatización, fenómeno por el cual el operador jurídico tiende a adoptar una postura pasiva de deferencia hacia la supuesta objetividad de la herramienta, perdiendo su capacidad crítica de análisis. Existe la amenaza real de que el juez o fiscal delegue su juicio por temor a ser investigado disciplinariamente si se aparta de un veredicto algorítmico que goza de un aura de infalibilidad técnica, convirtiendo la intervención humana en un trámite puramente simbólico. Para que el CHS sea efectivo, es imperativo que el operador no sea un humano decorativo, sino un agente moral con la competencia técnica y jurídica necesaria para invalidar la recomendación de la máquina cuando esta colisione con los valores constitucionales (Innerarity, 2024).

Esta problemática se agrava ante la opacidad intrínseca o fenómeno de la caja negra que caracterizó a PRiSMA y que como vimos encontramos en sistemas semejantes como el caso de COMPAS en Estados Unidos, cuyo método matemático de ponderación es ininteligible para los destinatarios de la decisión. Sin un nivel adecuado de transparencia sustantiva y explicabilidad, el derecho a la defensa se torna una garantía ilusoria, toda vez que el procesado se halla en un estado de indefensión epistémica al no poder conocer ni controvertir los fundamentos técnicos de su clasificación de riesgo. La asimetría informativa entre el Estado y el procesado impide que el juicio de peligrosidad estadística pueda ser

sometido a un escrutinio deliberativo, desplazando la racionalidad del derecho por una certeza computacional cerrada al debate.

Finalmente, el deber de motivación de las decisiones judiciales constituye el núcleo esencial del debido proceso, exigiendo que toda providencia sea el resultado de un ejercicio interpretativo calificado y no de una adhesión irreflexiva a un resultado automatizado. Si un funcionario judicial se apoya de un sistema de IA sin comprender su lógica interna, incurre en una motivación aparente, vulnerando el principio de razón suficiente que debe sustentar cualquier restricción a la libertad personal. Al respecto, la Corte Constitucional colombiana, en la sentencia T-323 de 2024, ha sido enfática al precisar que el uso de la inteligencia artificial no exime al servidor público de su responsabilidad en la verificación rigurosa de la información, consagrando el principio de no sustitución de la racionalidad humana como un límite infranqueable en la gestión de la justicia.

7.5 Caso Fiscal Watson: analítica investigativa y opacidad algoritmica

7.5.1. Igualdad material y no discriminación. La Automatización del Sesgo Histórico

En el Estado Social de Derecho, la igualdad material trasciende la mera equivalencia formal ante la ley para exigir condiciones reales y efectivas que proscriban toda forma de discriminación, sea directa, indirecta o sistémica (ACNUDH, 2022). En el caso de Fiscal Watson, esta garantía se ve tensionada por la calidad y naturaleza de los datos que alimentan el sistema. Fiscal Watson indexa y analiza más de 13 millones de noticias criminales alojadas en el Sistema Penal Oral Acusatorio (Palacios et al., 2024). No obstante, la persecución penal es, por naturaleza, una actividad que puede arrastrar sesgos históricos (Morales Higuita et

al., 2021). Al utilizar variables como la georreferenciación, el modus operandi y la reincidencia para asociar casos, el algoritmo corre el riesgo de operar sobre proxies de exclusión. Si los datos de entrada reflejan una selectividad policial histórica contra ciertos grupos o territorios vulnerables, el sistema simplemente automatiza y amplifica dichas asimetrías (Morales Higuita et al., 2021).

Bajo la lógica del fenómeno "garbage in, garbage out" (basura entra, basura sale), un dato de baja calidad o sesgado produce resultados que perpetúan trayectorias de criminalización de las que el individuo difícilmente puede escapar (Kusak, 2022, citado en Palacios et al., 2024, p. 15). Esto configura una discriminación indirecta, donde una práctica en apariencia neutra —el análisis estadístico masivo— produce efectos desproporcionadamente perjudiciales para poblaciones históricamente marginadas

7.5.2. Dignidad Humana

El uso de Fiscal Watson para sugerir vocación de éxito o priorizar investigaciones basadas en patrones de big data plantea el riesgo de reducir la singularidad del sujeto a una etiqueta de riesgo o a una asociación estadística. Cuando la administración de justicia privilegia la racionalidad objetiva del modelo sobre el contexto humano y social, la persona es instrumentalizada en favor de una métrica de eficiencia administrativa (Gómez et al., 2025).

Asimismo, la dignidad se ve comprometida por la opacidad del algoritmo. Al ser Fiscal Watson una herramienta desarrollada por un tercero privado (IBM) y protegida por derechos de autor, su funcionamiento interno permanece como una caja negra (Morales Higuita et al., 2021). Privar a un sujeto de conocer las razones exactas por las cuales fue

asociado a una cadena criminal o perfilado de cierta manera lesiona su autonomía y su derecho a ser tratado con respeto, pues se le sustrae la capacidad de controvertir racionalmente las premisas de la decisión estatal.

7.5.3. Control Humano Significativo

Como hemos visto con antelación, el Control Humano Significativo no debe entenderse como una supervisión meramente formal o decorativa donde el funcionario se limita a ratificar la salida del sistema, por el contrario, exige que el operador humano tenga la capacidad real de comprender, cuestionar y, si es necesario, apartarse del resultado algorítmico. En la implementación de Fiscal Watson, se advierte un riesgo de automatización de facto. Este fenómeno ocurre mediante el *heurístico de ajuste y anclaje* (Morales et al., 2021), donde el investigador toma el reporte de Watson como una verdad científica de base y ajusta su juicio a partir de ese valor inicial, otorgándole una presunción de veracidad injustificada por el solo hecho de ser un producto tecnológico.

Para que el control sea realmente significativo, la institución debe garantizar la trazabilidad y explicabilidad del proceso. Sin embargo, el Estado colombiano carece actualmente de métricas públicas de éxito o análisis de impacto en derechos específicos para esta herramienta (Palacios et al., 2024). La ausencia de una auditoría externa y la dependencia tecnológica del proveedor privado fragmentan la responsabilidad institucional y debilitan el control humano sobre el ciclo de vida del sistema

Así, la implementación de Fiscal Watson se inserta en una lógica de gubernamentalidad algorítmica, un modo de ejercicio del poder punitivo que busca gobernar a través de la predicción y segmentación de poblaciones basadas en datos (Innerarity, 2025).

Este sistema configura un nuevo panóptico digital donde la vigilancia no es directa, sino mediada por la indexación masiva de relatos y metadatos que hacen visible al ciudadano de forma permanente.

La tecnofascinación por resolver la congestión judicial mediante herramientas de analítica experta ha desplazado el debate sobre la soberanía de los datos y el impacto ético de estas tecnologías (Aguerre & Bustos, 2021, citado en Palacios et al., 2024). El control social se vuelve así más sutil pero omnipresente, en tanto la herramienta puede asociar investigaciones no solo en la etapa de indagación, sino incluso en análisis de contexto de justicia transicional, ampliando el radio de vigilancia sobre la esfera privada de los sujetos vinculados al proceso penal.

7.6 Síntesis reconstrucción de umbrales de validez constitucional

Tabla 9
Síntesis comparada de umbrales de validez constitucional

Garantía	Umbral (condición mínima)	Justificación constitucional (criterio)	Señales de incumplimiento
Igualdad material	El diseño y despliegue no puede concentrar cargas en poblaciones históricamente vulneradas sin justificación estricta	Evita que la neutralidad aparente consolide desigualdad material	Focalización opaca, perfilado territorial, cargas asimétricas.
No discriminación	Debe evitarse discriminación directa e indirecta por proxies (datos/modelo) y existir evaluación de impacto robusta y verificable	Impide efectos diferenciados injustificados y exige mitigación verificable	Proxies sensibles, datos sesgados, disparidades persistentes
Dignidad humana	El sistema no puede cosificar o estigmatizar mediante etiquetas, alertas	Preserva a la persona como fin y limita la gestión	Puntaje como etiqueta fija, automatización de

Garantía	Umbral (condición mínima)	Justificación constitucional (criterio)	Señales de incumplimiento
	o asociaciones de riesgo sin salvaguardas efectivas y control humano real.	punitiva por perfiles	sospecha, ausencia de recursos reales
Control humano significativo	Debe haber supervisión efectiva en el ciclo de vida del sistema (ex ante, durante y ex post), con posibilidad real de comprensión, apartamiento y registro.	Evita delegación encubierta y asegura responsabilidad institucional	Humano decorativo, automatización de facto, ausencia de registro.

8. CAPÍTULO IV. Evaluación garantista de los casos analizados

Esta Capítulo final representa la culminación del proceso investigativo, donde se contrasta la caracterización técnico-institucional (Capítulo II) con los umbrales de validez constitucional reconstruidos (Capítulo III).

Así las cosas, el objetivo específico es evaluar cómo la configuración técnico-institucional de los sistemas analizados incide sobre las garantías del Estado Social de Derecho previamente delimitadas en la investigación: igualdad material, no discriminación y dignidad humana, incorporando el control humano significativo como condición instrumental de exigibilidad y control. En coherencia con el diseño metodológico del trabajo, esta evaluación no reproduce la descripción empírica del caso, ya desarrollada en el Capítulo II ni reconstruye nuevamente los umbrales de validez constitucional, tarea cumplida en la Capítulo III, sino que aplica esos umbrales a los casos concretos para determinar el grado de cumplimiento y las desviaciones relevantes.

A través de esta evaluación, se determina en qué medida el despliegue de la Inteligencia Artificial (IA) en funciones estatales de control y vigilancia preserva o erosiona las garantías del Estado Social de Derecho

8.1 Caso COMPAS

La evaluación del caso COMPAS parte de tres premisas ya verificadas. Primero, que se trata de una herramienta de evaluación actuarial de riesgo y necesidades utilizada en distintas etapas del proceso penal y penitenciario estadounidense, cuya salida se integra en decisiones de clasificación, supervisión, manejo de casos y, en ciertos escenarios, informes pre-sentenciales. Segundo, que el sistema opera sobre una combinación de registros oficiales y entrevistas y/o autorreporte, incorporando variables que pueden funcionar como proxies de desigualdad estructural. Y tercero, que su uso se encuentra atravesado por problemas relevantes de opacidad, trazabilidad limitada, controversias empíricas sobre disparidades de error y un control humano condicionado por la propia arquitectura del sistema y por la presión institucional de eficiencia. Todo ello fue previamente descrito en la caracterización técnico-institucional del caso y constituye el soporte probatorio de la presente evaluación.

8.1.1. Igualdad material

El estándar aplicable en esta garantía exige que el diseño y despliegue de una herramienta algorítmica estatal no concentre cargas de vigilancia, restricción o encierro sobre poblaciones históricamente vulneradas sin una justificación constitucional estricta y sin salvaguardas verificables que permitan neutralizar la reproducción de desigualdades preexistentes. En otras palabras, el uso estatal de IA no puede reducir la igualdad a una neutralidad formal del cálculo si, en la práctica, la operación del sistema agrava o consolida

desigualdades materiales. Esta fue precisamente la función que se le asignó al umbral de igualdad material en la Tabla 2, y sobre ese parámetro debe leerse el caso COMPAS y los siguientes.

Aplicado a la evidencia reconstruida, COMPAS compromete este umbral por varias razones concurrentes. El sistema se aplica sobre poblaciones ya insertas en el circuito penal, es decir, sobre sujetos que previamente han sido alcanzados por estructuras estatales de vigilancia y castigo. Adicionalmente, se alimenta de registros oficiales que no son hechos neutrales, sino huellas institucionales producidas en contextos de policiamiento selectivo, criminalización diferencial y control intensificado sobre ciertos territorios y grupos. En este sentido, la lógica actuarial de COMPAS no parte de un terreno igualitario, sino de un universo de datos atravesado por desigualdades históricas que el sistema transforma en criterios operativos de riesgo (Mayson, 2018).

Esto se vuelve especialmente relevante cuando se observa que la herramienta no se limita a organizar información, sino que distribuye materialmente cargas y beneficios: intensifica vigilancia, restringe libertad, justifica modalidades de supervisión más severas o, en sentido inverso, habilita alternativas al encierro (Equivant, 2017). El problema constitucional surge cuando esa redistribución se apoya en información que refleja desigualdades pasadas y las traduce en riesgo futuro, sin un dispositivo robusto de corrección material. Así, la igualdad material se ve tensionada porque el sistema trata como equivalentes situaciones que no lo son, y lo hace bajo una racionalidad de eficiencia que invisibiliza el punto de partida desigual. En este caso, la aparente objetividad del cálculo no corrige la desigualdad; más bien la reordena y la proyecta hacia el futuro.

A ello se suma la evidencia empírica más influyente del caso, que reportó disparidades significativas en los tipos de error del sistema, especialmente en la clasificación de personas negras como de "alto riesgo" en proporciones superiores a las registradas para personas blancas (Angwin et al., 2016). Aunque esta evidencia fue objeto de controversia metodológica y de respuestas defensivas centradas en otras métricas (Mayson, 2018), lo cierto es que para el juicio de igualdad material la cuestión no se agota en si el sistema satisface o no una métrica estadística específica, sino en si produce cargas materialmente desiguales sobre grupos ya expuestos a mayor control penal. En esa medida, la evidencia de disparidades, sumada a la estructura de datos y al contexto de aplicación, permite concluir que COMPAS no satisface plenamente el umbral de igualdad material.

8.1.2. No discriminación

El estándar de no discriminación exige evitar tanto la discriminación directa como la discriminación indirecta producida por el uso de proxies, por el sesgo de los datos o por disparidades persistentes en los resultados. En el contexto de sistemas de IA estatal, este umbral no se satisface simplemente eliminando variables sensibles explícitas, como la raza, si el instrumento sigue operando mediante variables fuertemente correlacionadas con categorías protegidas o con posiciones estructurales de desventaja. Además, el estándar exige no solo identificar esos riesgos, sino contar con mecanismos verificables de evaluación y mitigación.

El caso COMPAS es paradigmático precisamente porque pone de relieve este problema. Aunque el sistema se presenta como formalmente ciego a la raza, utiliza variables y registros que funcionan como proxies de pertenencia racial, posición socioeconómica y

territorialidad: historial penal, educación, empleo, estabilidad residencial, entre otros (Mayson, 2018). La propia evidencia técnica y doctrinal reconstruida en el Capítulo II muestra que la omisión de la raza como input no elimina la posibilidad de discriminación, cuando la estructura correlacional de los datos continúa reproduciendo las asimetrías de la vigilancia y del castigo. Esto significa que la discriminación indirecta no se introduce a pesar del diseño, sino a través de la forma en que el sistema traduce una historia social desigual en criterios de clasificación.

La evidencia documentada sobre disparidades en falsos positivos y falsos negativos por subgrupos raciales refuerza esa conclusión. El debate entre ProPublica y Equivant mostró que el desacuerdo no se ubica tanto en si existe una diferencia estadística, sino en qué métrica debe considerarse jurídicamente relevante (Angwin et al., 2016; Mayson, 2018). Pero desde una perspectiva constitucional, esa discusión no puede resolverse solo al nivel de la técnica, porque el problema no es exclusivamente si el sistema está "bien calibrado", sino si las diferencias en error y clasificación terminan generando efectos diferenciados injustificados sobre personas pertenecientes a grupos históricamente marginados.

Por ello, en COMPAS la no discriminación aparece comprometida no tanto porque el sistema mencione explícitamente criterios sospechosos, sino porque incorpora una infraestructura de datos y una lógica de inferencia que proyectan desventajas previas hacia decisiones futuras. El carácter opaco del modelo y la ausencia de una evaluación pública, homogénea y exigible sobre su impacto diferencial refuerzan esta conclusión.

8.1.3. Dignidad humana

La dignidad humana funciona aquí como límite material frente a la reducción de la persona a una etiqueta de riesgo. El estándar aplicable exige que el sistema no cosifique ni estigmatice al sujeto mediante clasificaciones que sustituyan la valoración individual por una probabilidad estadística, especialmente cuando esas clasificaciones pueden afectar libertad, tratamiento institucional o intensidad de control. No se trata de prohibir cualquier forma de organización de información, sino de impedir que la persona sea tratada como mero objeto de gestión actuarial.

En COMPAS, la afectación a la dignidad se manifiesta en la propia naturaleza del output: el individuo es situado en un decil y en una categoría de riesgo construida a partir de patrones observados en grupos similares, no de una certeza individual sobre su conducta (Equivant, 2017). Aunque la documentación técnica y la jurisprudencia insisten en que el puntaje no debe leerse como una predicción individual concluyente, en la práctica el puntaje opera como una etiqueta estabilizadora de la identidad procesal del sujeto (Harvard Law Review, 2017). La persona deja de ser considerada prioritariamente a partir de su singularidad y pasa a ser leída como miembro de una categoría estadística, susceptible de una administración diferenciada según su riesgo.

Esa lógica resulta especialmente problemática cuando se combina con el carácter opaco del sistema y con la presión institucional por decisiones rápidas y consistentes (Washington, 2019). En tales condiciones, el riesgo deja de ser un dato auxiliar y se convierte en una forma de objetivación del individuo. El tránsito desde “persona” hacia “perfil” erosiona la idea de que cada caso requiere una valoración individual, situada y abierta a contradicción. Desde esta perspectiva, la dignidad no se ve afectada solo por un eventual

error, sino por la estructura misma de una decisión que reduce al sujeto a una puntuación y que organiza la respuesta institucional a partir de esa puntuación.

Ahora bien, a diferencia de igualdad material y no discriminación, aquí sí existen algunos elementos de contención parcial. El propio caso Loomis (2016) reconoce expresamente que el puntaje no puede usarse como fundamento determinante y que su alcance inferencial es grupal, no individual. Esa advertencia judicial limita, al menos en el plano formal, el riesgo de cosificación total. Sin embargo, esa contención no elimina la afectación, porque no suprime la carga simbólica y práctica de la etiqueta de riesgo dentro del proceso decisional.

8.1.4. Control humano significativo

De acuerdo con la metodología de esta tesis, el control humano significativo opera como una garantía instrumental: no constituye por sí mismo una garantía sustantiva independiente, pero hace exigibles y controlables la igualdad, la no discriminación y la dignidad. El estándar aplicable exige, por tanto, una intervención humana efectiva en las fases de diseño, uso y revisión del sistema, así como una capacidad real de apartarse del output y de motivar tal apartamiento.

En el caso COMPAS, el control humano existe formalmente en varios puntos del ciclo de vida del sistema. Interviene en la definición institucional del uso, en la captura y corrección de datos, en la lectura del reporte y en la decisión final. Además, el sistema prevé mecanismos como el override, y la jurisprudencia insiste en que el puntaje no debe ser determinante ni sustituir el juicio del decisor. Desde esta perspectiva, no se trata de un sistema de automatización total (Equivant, 2017; State v. Loomis, 2016).

Sin embargo, el problema no radica en la ausencia absoluta de un humano, sino en la calidad de esa intervención. Cuando el sistema funciona como caja negra y el operador no puede comprender suficientemente la lógica de transformación entre datos y puntaje, la posibilidad de apartarse se reduce a una facultad más formal que sustantiva. A ello se suma el riesgo de anclaje cognitivo: la aparente experticia del número y la presión por eficiencia administrativa pueden llevar al juez o al funcionario a reproducir el output sin escrutinio material, aun cuando en teoría conserve la última palabra (Harvard Law Review, 2017; Washington, 2019). En tal contexto, el control humano deja de ser plenamente significativo y se aproxima a una forma de validación decorativa del resultado algorítmico.

Por lo anterior, aunque COMPAS no elimina al decisor humano, sí lo coloca en una posición de dependencia epistémica frente al puntaje, especialmente cuando no existen condiciones adecuadas de explicabilidad, trazabilidad y discusión crítica del resultado. El control humano, entonces, aparece debilitado por la propia arquitectura del sistema.

8.1.5. Conclusión evaluativa del caso COMPAS

El caso COMPAS permite concluir que su incidencia sobre las garantías del Estado Social de Derecho es críticamente problemática. El sistema no cumple el umbral de no discriminación, debido al uso de proxies estructurales y a la evidencia de disparidades diferenciales en los resultados, sin contar con un esquema suficiente y públicamente verificable de mitigación. Asimismo, cumple parcialmente los estándares de igualdad material, dignidad humana y control humano significativo, pues, aunque existen algunas salvaguardas formales, particularmente advertencias judiciales y la prohibición de usar el puntaje como fundamento único, esas salvaguardas resultan insuficientes para neutralizar los

efectos de una infraestructura de datos sesgada, de una lógica de clasificación actuarial opaca y de un uso institucional fuertemente condicionado por la eficiencia.

En consecuencia, COMPAS muestra cómo una herramienta algorítmica puede presentarse como apoyo técnico sin dejar de reconfigurar materialmente la justicia penal: redistribuye cargas sobre poblaciones vulnerables, produce riesgos de discriminación indirecta, reduce a la persona a un perfil estadístico y debilita la capacidad del control humano de operar como garantía efectiva. Desde el punto de vista de esta investigación, el caso constituye un estándar de contraste especialmente valioso para el contexto colombiano, porque anticipa los riesgos constitucionales de importar o naturalizar lógicas actuariales de perfilamiento sin someterlas a exigencias estrictas de transparencia, control humano significativo, contradicción real y corrección material de desigualdades.

8.2 Caso SyRI

En la caracterización técnico-institucional quedó establecido que SyRI se trató de un sistema estatal de detección de fraude en el ámbito del bienestar social de Países Bajos, apoyado en el cruce interinstitucional de datos administrativos, aplicado mediante focalización territorial en barrios vulnerables, con un modelo de perfilamiento opaco, una salida consistente en alertas de riesgo y una gobernanza marcada por déficits de transparencia, trazabilidad y control humano materialmente suficiente. También quedó demostrado que su principal impacto documentado fue jurisdiccional: la decisión del tribunal de La Haya que concluyó que el marco SyRI era incompatible con el artículo 8 del Convenio Europeo de Derechos Humanos, por ausencia de salvaguardas adecuadas y por su potencial de producir estigmatización y discriminación indirecta.

Sobre esa base, la evaluación garantista del caso permite sostener que SyRI no fue un simple instrumento técnico de optimización administrativa, sino un dispositivo de perfilamiento estatal que, al operar sobre datos históricos, focalización territorial y reglas de inferencia no auditables, reconfiguró de forma intensa la relación entre vigilancia pública y garantías constitucionales. Lo que aquí corresponde, por tanto, es determinar en qué medida esa configuración resulta compatible o incompatible con los umbrales ya fijados.

8.2.1. Igualdad material

El estándar aplicable exige que el diseño y despliegue del sistema no concentre cargas de vigilancia y sospecha sobre poblaciones históricamente vulneradas sin una justificación estricta, verificable y constitucionalmente suficiente. En otras palabras, la igualdad material no se satisface cuando el Estado trata como neutral una arquitectura de control que recae selectivamente sobre quienes ya ocupan posiciones estructurales de desventaja.

En SyRI, la evidencia del caso muestra que el sistema no se aplicó de manera general y neutra sobre la población, sino que fue desplegado en vecindarios previamente identificados como “problemáticos”, caracterizados por altos índices de pobreza, desempleo, beneficiarios de asistencia y presencia de población migrante o perteneciente a minorías étnicas. En términos prácticos, el sistema no solo detectaba riesgo, sino que redistribuía el umbral de sospecha hacia sectores específicos del espacio social, concentrando allí los recursos de fiscalización y control administrativo. La selección territorial no fue un dato accesorio del diseño, sino uno de sus rasgos constitutivos más relevantes (Rechtbank Den Haag, 2020).

Desde la perspectiva garantista, esto compromete de manera directa la igualdad material. El problema no consiste únicamente en que el sistema se use más en ciertos lugares, sino en que convierte la vulnerabilidad socioeconómica en condición de exposición reforzada al control estatal. De esta forma, la pobreza deja de ser una situación que activa deberes de protección del Estado y pasa a funcionar, indirectamente, como un indicador de riesgo administrativo. La consecuencia es que el sistema no corrige desigualdades previas, sino que las toma como punto de partida para redistribuir la vigilancia y, con ello, intensifica la carga de sospecha precisamente sobre quienes ya se encuentran en una posición estructuralmente débil (Bekker, 2021). Ese patrón resulta incompatible con un entendimiento material de la igualdad, en el cual el Estado no puede organizar su poder de control partiendo de categorías territoriales o sociales que repliquen exclusión.

A ello se suma que el propio marco normativo y operativo del sistema no ofrecía salvaguardas suficientemente robustas para neutralizar ese sesgo estructural. No existía una arquitectura verificable que permitiera demostrar que la focalización territorial respondía a criterios estrictamente necesarios, proporcionados y no discriminatorios. Por el contrario, la opacidad del modelo y la baja explicabilidad impedían controlar si la concentración territorial respondía a una decisión justificada o si operaba, más bien, como una forma de vigilancia de clase automatizada (Rechtbank Den Haag, 2020). En este punto, la igualdad material se ve erosionada no solo por el resultado, sino por la imposibilidad de someter el diseño a un escrutinio público suficiente.

8.2.2. No discriminación

El estándar aplicable exige que el sistema evite discriminación directa e indirecta, no sólo en el plano formal de sus variables, sino en la estructura de sus resultados y en los efectos que produce sobre grupos protegidos. En el contexto algorítmico, ello significa que la omisión de categorías sensibles explícitas no basta cuando el modelo opera mediante proxies que reproducen diferencias injustificadas entre grupos.

La evidencia técnico-institucional de SyRI muestra que el sistema trabajó con un universo amplio de datos administrativos —vinculados a ingresos, vivienda, deudas, educación, salud, beneficios y otros registros estatales— cuya agregación permitía construir perfiles de riesgo sobre personas y hogares. Aunque el régimen no se presentaba como un sistema que utilizara categorías sensibles en forma directa, la propia configuración de sus datos y su despliegue territorial hacían posible que variables aparentemente neutras operaran como indicadores indirectos de condición socioeconómica, origen migratorio o pertenencia territorial, proyectando hacia el sistema las desigualdades históricas ya presentes en las bases administrativas y en las prácticas de fiscalización (van Bekkum & Zuiderveen, 2021).

Desde esta perspectiva, la afectación a la no discriminación no depende de demostrar que SyRI preguntaba por la raza o por la condición migratoria, sino de constatar que su forma de operar convertía en criterio de sospecha elementos estrechamente ligados a posiciones estructurales de marginalidad. La combinación entre focalización en barrios vulnerables, uso masivo de datos administrativos y opacidad en la inferencia del riesgo configura una forma típica de discriminación indirecta: el sistema no selecciona a ciertos grupos nombrándolos, pero sí a través de la red de variables y contextos que los representan funcionalmente. En esa medida, el sesgo no es accidental ni meramente técnico, sino estructural.

La gravedad de esta afectación se incrementa por la ausencia de mecanismos de control suficientes. En el caso no hubo una evaluación de impacto en derechos capaz de demostrar, de manera pública y verificable, que el sistema había controlado o mitigado adecuadamente el riesgo de discriminación indirecta. Por el contrario, el tribunal de La Haya consideró que la opacidad y la falta de salvaguardas impedían asegurar que el sistema no produjera precisamente ese tipo de efectos. De este modo, el sistema no sólo estaba expuesto al riesgo de discriminación, sino que lo hacía sin ofrecer las garantías necesarias para probar lo contrario.

8.2.3. Dignidad humana

La dignidad humana exige que la persona sea tratada como un fin en sí mismo y no como mero objeto de administración, clasificación o sospecha. En el contexto de sistemas algorítmicos estatales, este estándar impone un límite material a la reducción del individuo a una etiqueta de riesgo producida por datos históricos y modelos opacos, especialmente cuando esa etiqueta activa investigaciones o condiciona la relación del sujeto con la administración.

En SyRI, la afectación a la dignidad se manifiesta en la propia lógica del sistema. El individuo no era valorado en su singularidad, sino procesado como parte de un agregado de datos que, una vez cruzados, podían transformarse en una alerta de riesgo. Esa alerta no era una sanción definitiva, pero sí bastaba para situar a la persona dentro de un circuito reforzado de vigilancia administrativa. El sujeto aparecía así reducido a un perfil probabilístico, construido no desde una conducta probada individualmente, sino desde patrones inferidos a

partir de datos pasados, contextos de residencia y categorías administrativas heterogéneas (Rechtbank Den Haag, 2020).

Esta dinámica compromete la dignidad porque sustituye la interlocución entre persona y Estado por una relación en la que la administración observa, perfila y selecciona sin ofrecer una explicación suficiente del porqué. La opacidad del modelo, sumada a la inexistencia de una explicabilidad robusta para el afectado, produce una forma de humillación administrativa silenciosa en donde la persona puede verse marcada por el sistema sin contar con condiciones reales para comprender, anticipar o discutir la lógica que motivó su selección. En este punto, la dignidad no se afecta solo por el posible error, sino por la estructura misma de un procedimiento que trata al individuo como objeto de gestión de riesgo y no como sujeto de derechos capaz de confrontar el fundamento de la sospecha (Bekker, 2021; Rachovitsa & Johann, 2022).

Además, el despliegue en "barrios problema" introduce una dimensión adicional de estigmatización. La persona no solo es tratada como riesgo en razón de sus datos, sino también en razón del lugar social y territorial que ocupa. El sujeto aparece así doblemente reducido: por su perfil estadístico y por su inscripción territorial en una cartografía administrativa de la sospecha. Esta forma de cosificación es incompatible con la dignidad humana, en la medida en que convierte al ciudadano en objeto de una gestión preventiva de la pobreza y el fraude, subordinando su condición de fin en sí mismo a una racionalidad de eficiencia fiscal y control.

8.2.4. Control humano significativo

En el diseño metodológico de esta tesis, el control humano significativo opera como una condición instrumental indispensable para hacer exigibles y controlables las garantías sustantivas de igualdad material, no discriminación y dignidad humana. El estándar aplicable exige una supervisión humana efectiva en el ciclo de vida del sistema ex ante, durante la operación y ex post, y, sobre todo, una capacidad real de comprender, validar, cuestionar y, si es necesario, apartarse del output algorítmico.

En SyRI existían formalmente intervenciones humanas en distintas fases (Rechtbank Den Haag, 2020), había una decisión inicial de activar el sistema mediante proyectos interinstitucionales, una revisión ministerial de requisitos, una operación en dos fases con intervención de unidades técnicas y analíticas, y una retroalimentación ex post que permitía ajustar o evaluar resultados. Desde una mirada superficial, ello podría sugerir la existencia de control humano. Sin embargo, el criterio relevante no es la mera presencia de funcionarios en la cadena, sino el carácter significativo de su intervención. Y es precisamente allí donde el caso revela su principal fractura.

La intervención humana se encontraba estructuralmente debilitada por la falta de transparencia del modelo. Si quienes participan en la operación o revisión no pueden reconstruir ni explicar con claridad la lógica que conduce a una alerta de riesgo, entonces su intervención se reduce a administrar resultados ya filtrados por una caja negra. En ese contexto, el humano deja de ser un verdadero decisor y se convierte en un operador de salidas cuyo fundamento no controla. A su vez, la posibilidad de apartarse del output pierde densidad práctica cuando no existe una base comprensible y verificable para revisar críticamente el resultado (Rechtbank Den Haag, 2020).

Este problema se agrava desde la perspectiva del afectado. El sistema contemplaba la posibilidad de solicitar información sobre la inclusión en el registro, pero también permitía rechazar ese acceso cuando hubiese investigación en curso, precisamente para proteger el modus operandi del sistema. Ello significa que el supuesto control humano no se traducía en una capacidad efectiva de defensa o de contradicción para el sujeto, sino que operaba dentro de una arquitectura de asimetría informacional. De ahí que la participación humana, aun existente formalmente, no satisfaga el estándar de control significativo, se trataba de una intervención funcionalmente subordinada a un modelo opaco y escasamente auditable.

8.2.5. Conclusión evaluativa del caso SyRI

Aplicados los umbrales reconstruidos en la Capítulo III a la evidencia técnico-institucional sistematizada en la Capítulo II, el caso SyRI permite concluir que su configuración resulta incompatible con las garantías del Estado Social de Derecho que esta investigación toma como eje de análisis. El sistema no cumple los estándares de igualdad material, no discriminación y dignidad humana, y tampoco satisface la exigencia de control humano significativo como condición instrumental de legitimidad.

La razón central de este resultado es que SyRI articuló cuatro elementos particularmente lesivos: focalización territorial de la vigilancia, uso masivo de datos administrativos con fuerte capacidad de proxy, opacidad sustantiva del modelo y déficits de control humano real. Esa combinación permitió trasladar al ámbito algorítmico una lógica de sospecha estructural sobre sectores vulnerables, sin ofrecer garantías suficientes de explicabilidad, contradicción y justificación estricta. El caso muestra, por tanto, que cuando el Estado utiliza sistemas de perfilamiento en contextos de bienestar social sin salvaguardas

robustas, la eficiencia administrativa puede transformarse en una forma tecnificada de exclusión.

Desde el punto de vista comparado, SyRI constituye una advertencia especialmente fuerte para la investigación puesto que demuestra que la ausencia de transparencia, trazabilidad y control humano no es una deficiencia meramente procedimental, sino una falla que puede erosionar de manera estructural la igualdad material, la no discriminación y la dignidad humana. En esa medida, el caso ofrece un estándar de contraste decisivo para la evaluación de los demás sistemas examinados en esta tesis y confirma que la validez constitucional de la IA estatal depende no solo de lo que el sistema pretende hacer, sino de cómo está diseñado, sobre qué datos opera, a quiénes expone y qué capacidad real de control conserva el ser humano frente al resultado algorítmico.

8.3 Caso AFR Locate

Como quedó demostrado en el Capítulo II, AFR Locate se configura como un sistema de reconocimiento facial en vivo desplegado por la policía de Gales del Sur en espacios públicos, que operó mediante captura masiva de rostros, comparación biométrica frente a una watchlist, generación de una alerta de coincidencia y verificación humana posterior. También se estableció que su despliegue se apoya en documentos de gobernanza como políticas internas y evaluaciones de impacto, pero conserva una opacidad relevante respecto del funcionamiento del algoritmo propietario. A ello se suma una evidencia empírica de falsos positivos significativos y un control judicial que, en el caso Bridges, identificó déficits de legalidad, discrecionalidad excesiva y fallas en el cumplimiento del deber de igualdad del sector público.

Sobre esa base, la evaluación garantista permite sostener que AFR Locate no es un simple instrumento auxiliar de identificación, sino una forma tecnológicamente sofisticada de vigilancia estatal en espacio público que reconfigura la relación entre anonimato, libertad de circulación, sospecha policial y control administrativo. Lo que corresponde, entonces, es establecer en qué medida esa configuración satisface o vulnera los umbrales reconstruidos en la Capítulo III.

8.3.1. Igualdad material

El estándar aplicable en esta garantía exige que el diseño y despliegue de una herramienta algorítmica estatal no concentre cargas de vigilancia sobre grupos o espacios históricamente vulnerados sin criterios estrictos, verificables y compatibles con el Estado Social de Derecho. La igualdad material no se satisface cuando el uso de la tecnología, bajo la apariencia de neutralidad operativa, intensifica la exposición al control precisamente sobre sectores o territorios ya sometidos a mayor escrutinio estatal.

En AFR Locate, este estándar se ve comprometido porque el sistema no opera de forma difusa y universal, sino mediante despliegues situacionales en lugares previamente seleccionados por la autoridad policial. La propia reconstrucción del caso muestra que el territorio impactado se delimita por decisiones operativas sobre dónde desplegar las cámaras y a quiénes incorporar en las watchlists, lo cual otorga a la autoridad un margen significativo de selección. En el litigio Bridges, este punto fue especialmente sensible: la Corte de Apelación cuestionó que el marco normativo no delimitara con suficiente densidad el quién y el dónde de la vigilancia, abriendo espacio a discrecionalidad excesiva en la selección de personas y lugares. Esa indeterminación es incompatible con un estándar de igualdad

material cuando la vigilancia algorítmica se proyecta en entornos urbanos concretos y sobre el flujo general de transeúntes (R (Bridges) v Chief Constable of South Wales Police & Ors, 2020).

El problema se agrava porque la propia lógica de despliegue puede operar mediante proxies territoriales y sociales. Si la selección de watchlists y de zonas de uso se articula en torno a objetivos policiales definidos por patrones históricos de vigilancia, la herramienta deja de ser un simple apoyo técnico y se convierte en un mecanismo que puede predeterminar la exposición diferencial al control. En tal escenario, el sistema no corrige desigualdades previas, sino que las reorganiza en clave tecnológica, proyectando sobre ciertos sectores del espacio público una vigilancia intensificada. La igualdad material se ve así erosionada no solo por el resultado final, sino por la estructura de selección que antecede la propia activación del sistema.

8.3.2. No discriminación

El estándar de no discriminación exige evitar no solo la discriminación directa, sino también la discriminación indirecta derivada de reglas aparentemente neutras que, en la práctica, producen afectaciones desproporcionadas sobre grupos protegidos. En el contexto de tecnologías biométricas, ello implica evaluar tanto los criterios de selección de las watchlists y de los lugares de despliegue como las diferencias de desempeño del sistema por sexo, raza u otros rasgos demográficos. Además, el estándar exige que la autoridad investigue y mitigue esos riesgos de forma independiente y verificable.

En AFR Locate, la evidencia del caso es especialmente fuerte en este punto. La Corte de Apelación determinó que la policía de Gales del Sur no cumplió con el Deber de Igualad

en el Sector Publico al no verificar de manera suficientemente independiente si el software presentaba sesgos por raza o sexo (R (Bridges) v Chief Constable of South Wales Police & Ors, 2020). Esta conclusión es central, porque muestra que el problema no se reduce a una discusión teórica sobre sesgo algorítmico, sino que se materializa en un incumplimiento institucional del deber de examinar el impacto diferencial de la herramienta. A ello se suma la evidencia técnica general según la cual los sistemas de reconocimiento facial tienden a mostrar tasas de falsos positivos desproporcionadas en mujeres y minorías étnicas, lo que vuelve metodológicamente insostenible asumir neutralidad a partir de la mera ausencia de variables sensibles explícitas (Purshouse, 2022).

La discriminación indirecta se agrava por dos factores adicionales. Primero, la composición de las watchlists puede operar como proxy de condiciones sociales y territoriales, amplificando patrones previos de vigilancia selectiva. Segundo, la propia opacidad del algoritmo propietario limita la posibilidad de que los afectados o terceros independientes verifiquen si la herramienta funciona con igual desempeño entre subgrupos o si, por el contrario, reproduce asimetrías históricas. En esta medida, AFR Locate no solo expone a una afectación diferenciada, sino que lo hace sin un esquema de mitigación pública suficientemente robusto. Desde una lectura garantista, ello compromete directamente la prohibición de discriminación.

8.3.3. Dignidad humana

La dignidad humana exige que la persona no sea reducida a objeto de gestión, identificación o sospecha por parte del poder estatal. En el contexto del reconocimiento facial en vivo, este estándar se concreta en la prohibición de convertir el rostro y la presencia en el

espacio público en una estructura de información procesable sin salvaguardas materiales suficientes. No se trata solo de proteger datos biométricos en abstracto, sino de impedir que la tecnología transforme a la ciudadanía en una masa permanentemente disponible para clasificación y vigilancia.

AFR Locate compromete este umbral porque se basa en la captura masiva y el procesamiento biométrico de transeúntes que no han sido previamente individualizados en el terreno. El sistema no se limita a observar a quienes están en la watchlist; para poder localizarlos, escanea a todas las personas que transitan por el área de despliegue. Bajo esa lógica, el ciudadano común deja de ser simplemente un sujeto que circula en el espacio público y pasa a ser un cuerpo continuamente traducible en una firma biométrica comparable. Esa transformación del cuerpo en dato disponible para vigilancia constituye una forma de instrumentalización que afecta la dignidad, porque subordina la presencia física de la persona a una racionalidad de control preventivo (Purshouse, 2022).

La afectación se intensifica por la opacidad del sistema. El afectado no puede comprender con plenitud por qué fue señalado, qué rasgos fueron determinantes ni cómo se alcanzó el possible match. Aunque existe verificación humana posterior, ello no restituye plenamente la autonomía cuando el individuo ya ha sido sometido a una lógica de identificación algorítmica que lo clasifica sin una explicación inteligible (Davies et al., 2018).

La persona aparece, así como una tarjeta de identidad caminante, su cuerpo deja de ser presupuesto de libertad y se convierte en insumo de una gestión automatizada de seguridad. Con todo, a diferencia de lo que ocurre en no discriminación, aquí deben reconocerse algunas contenciones parciales: el sistema es usado de forma pública, la alerta

no equivale a identificación definitiva y existen reglas de borrado relativamente estrictas para no coincidencias. Esas medidas no eliminan la lesión potencial, pero sí impiden afirmar una cosificación absoluta en todos los casos.

8.3.4. Control humano significativo

En AFR Locate existe, formalmente, un esquema *de human-in-the-loop*. El despliegue es autorizado ex ante por personal responsable; la watchlist se construye y controla antes del operativo; durante la operación, cada alerta de posible coincidencia debe ser revisada por un operador humano antes de cualquier intervención; y ex post existen registros, evaluaciones internas y control judicial. Estos elementos muestran que no se trata de un sistema de automatización total. Además, en teoría, el operador puede apartarse de la alerta cuando la verificación visual o el contexto no justifican actuar (Surveillance Camera Commissioner, 2019).

Sin embargo, el problema garantista no es la ausencia de humano, sino el carácter materialmente insuficiente de su intervención. La propia lógica del sistema y el entorno operativo pueden producir un sesgo de automatización, la alerta se integra al flujo policial como una señal técnica que empuja la intervención y funciona como anclaje decisional, especialmente bajo presión temporal y confianza en la máquina (Fussey & Murray, 2019). A ello se añade la opacidad del algoritmo, si el operador no puede acceder a la lógica fina del match ni a los parámetros que justifican la coincidencia, su control se limita a revisar una imagen o un resultado, pero no el proceso que lo generó. En tales condiciones, el control humano corre el riesgo de convertirse en una coartada simbólica que reviste de legalidad

humana una decisión previamente condicionada por una arquitectura opaca (Davies et al., 2018).

Con todo, a diferencia de SyRI, en AFR Locate sí existen espacios procedimentales de apartamiento y una obligación de verificación antes de actuar. Eso impide hablar de una automatización completa del poder policial. Sin embargo, tampoco permite afirmar que el control sea plenamente significativo, porque el operador interviene sobre un resultado que no puede comprender en profundidad y bajo una dinámica institucional donde la alerta puede anclar la decisión. Por ello, el estándar solo se satisface parcialmente.

8.3.5. Conclusión evaluativa del caso AFR Locate

Aplicados los umbrales reconstruidos en la Capítulo III a la evidencia sistematizada en la Capítulo II, AFR Locate aparece como un caso de tensión garantista alta. El sistema no cumple los estándares de igualdad material y no discriminación, debido a la combinación de discrecionalidad excesiva en el despliegue, posible concentración desigual de vigilancia, insuficiencia en la evaluación independiente de sesgos y evidencia técnica de diferencias de desempeño entre grupos demográficos. Asimismo, cumple parcialmente los estándares de dignidad humana y control humano significativo: existen contenciones formales relevantes tales como verificación humana y reglas de borrado, pero estas no neutralizan completamente el carácter intrusivo de la captura biométrica masiva ni el riesgo de automatización de facto de la intervención policial.

El caso demuestra que, incluso cuando una tecnología no decide jurídicamente por sí sola, puede reconfigurar de forma intensa las garantías del Estado Social de Derecho si redistribuye la vigilancia en el espacio público, produce errores con impacto desigual, opera

bajo lógicas parcialmente opacas y se inserta en marcos normativos con discrecionalidad excesiva. AFR Locate ofrece así una advertencia constitucional clara, en tecnologías de vigilancia biométrica, la mera existencia de documentos de gobernanza o de una verificación humana posterior no basta para asegurar legitimidad. Se requieren criterios estrictos, verificables y públicamente controlables sobre quién es vigilado, dónde, por qué, con qué datos, con qué márgenes de error y bajo qué posibilidades reales de control y contradicción. Solo bajo esas condiciones podría sostenerse que una tecnología semejante no erosiona las garantías que el Estado Social de Derecho está llamado a proteger.

8.4 Caso PRisMA

En el caso de PRiSMA, la base fáctica relevante ya quedó establecida en la caracterización técnico-institucional. Se trata de una herramienta desarrollada por la Fiscalía General de la Nación para apoyar la solicitud de medidas de aseguramiento, a partir de una estimación del riesgo de reincidencia construida con registros del SPOA, SIEDCO, Policía e INPEC; su salida se expresa en un reporte probabilístico descargable en PDF, pensado para orientar la argumentación del fiscal, con un despliegue verificable al menos en un plan piloto en varias seccionales, bajo una gobernanza intraestatal y con transparencia parcial sobre la arquitectura del modelo. La evidencia pública, además, muestra limitaciones importantes en trazabilidad, explicabilidad y evaluación externa de impactos, así como incertidumbre sobre su continuidad en uso.

Sobre esa base, la evaluación garantista permite sostener que PRiSMA no es un simple mecanismo auxiliar de organización de información, sino una tecnología de predicción penal que se inserta en una de las decisiones más intensas del proceso penal: la restricción preventiva de la libertad. En consecuencia, la pregunta no es solo si “mejora” la

eficiencia institucional, sino si su diseño, sus datos, su opacidad y su rol práctico son compatibles con las exigencias de igualdad material, no discriminación, dignidad humana y control humano significativo que un Estado Social de Derecho impone.

8.4.1. Igualdad material

En PRiSMA, este umbral se ve comprometido porque la herramienta opera sobre personas ya insertas en el circuito penal colombiano y lo hace a partir de bases administrativas y judiciales históricas que incorporan trazas de capturas, imputaciones, antecedentes y trayectorias penitenciarias.

La propia reconstrucción del caso deja claro que esos insumos pueden contener huellas institucionales de vigilancia selectiva, prácticas policiales diferenciadas y desigualdades territoriales, por lo que el dato no puede asumirse como un reflejo neutral del riesgo, sino como una sedimentación de relaciones previas de poder punitivo. Además, la evidencia pública muestra que el despliegue verificable del sistema se concentró, al menos inicialmente, en seccionales y URI ubicadas en grandes centros urbanos como Bogotá y Medellín, donde la intensidad de capturas e imputaciones es mayor, lo que refuerza el riesgo de que la herramienta opere con mayor frecuencia precisamente donde la selectividad penal ya es más intensa (Gutiérrez & Muñoz, 2023; Fair Trials, 2024)

El problema garantista, entonces, no radica únicamente en que PRiSMA calcule un riesgo, sino en que ese cálculo puede servir para racionalizar la distribución de medidas intramurales usando como base trayectorias previamente construidas por un sistema penal desigual. Si la promesa de eficiencia consiste en concentrar la detención preventiva en quienes exhiben mayores niveles objetivos de riesgo, pero esos niveles se derivan de registros

históricamente atravesados por sesgos institucionales, el sistema corre el riesgo de convertir la desigualdad estructural en fundamento técnico para reforzar la restricción de libertad. En lugar de corregir desigualdades, las administra y las proyecta.

8.4.2. No discriminación

En PRiSMA, la afectación a esta garantía se configura precisamente un nivel indirecto. Los documentos públicos reconstruidos muestran que la herramienta se alimenta de registros de Policía, Fiscalía e INPEC, y que sus variables incluyen antecedentes de capturas, delitos previos, medidas previas e información penitenciaria (Torres et al,. 2022). Incluso cuando se sostenga que el sistema no incorpora variables sociodemográficas de forma explícita, en la caracterización del caso se logró reconocer que estos insumos pueden operar como proxies de pobreza, territorialidad y exposición desigual al policiamiento, trasladando al modelo las disparidades del sistema penal de origen (Fair Trials, 2024).

A ello se suma que la documentación pública no permite verificar con rigor un diccionario completo de variables, reglas de depuración, exclusión de variables sensibles o mecanismos robustos de mitigación de sesgos, lo que limita la posibilidad de descartar afectaciones discriminatorias de manera controlable (Díaz et al,. 2024).

Desde esta perspectiva, la discriminación indirecta no aparece como una hipótesis abstracta, sino como un riesgo estructural del modo en que el sistema traduce historia penal en riesgo futuro. Si la vigilancia y las capturas han recaído de manera más intensa sobre ciertos territorios o grupos sociales, y si esos registros se convierten luego en predictores relevantes de reincidencia, el algoritmo no descubre un riesgo preexistente: lo reorganiza y legitima en clave técnica.

8.4.3. Dignidad humana

La dignidad humana exige que el individuo sea tratado como fin en sí mismo y no como simple objeto de administración del riesgo. En el campo de la justicia predictiva, este estándar se concreta en la prohibición de reducir la singularidad biográfica y jurídica de la persona a una estimación probabilística que condicione su acceso a la libertad antes de una condena. La persona no puede convertirse en una proyección estadística del pasado sin que ello comprometa el núcleo de su autonomía y de su presunción de inocencia material.

En PRiSMA, la tensión con la dignidad es especialmente intensa porque el sistema se inserta en el momento procesal de solicitud de medida de aseguramiento, es decir, mucho antes de una sentencia definitiava, cuando la libertad del imputado se encuentra más estrechamente conectada con el estándar de excepcionalidad de la detención preventiva (Fair Trials, 2024). El output de la herramienta se materializa en un reporte que presenta un nivel de riesgo y orientaciones para el fiscal, pensado para respaldar la solicitud de una medida. Aunque el resultado se presenta como una probabilidad y no como una constatación, la propia estructura del sistema —un puntaje, niveles de riesgo, reportes de fácil circulación en audiencia— tiende a convertir al individuo en un perfil actuarial cuya trayectoria futura se vuelve administrable a partir de datos pasados (Torres et al,. 2022). La persona deja de ser vista primordialmente desde la singularidad del caso y pasa a ser leída como un riesgo calculado.

La afectación a la dignidad se agrava por dos elementos. Primero, por la opacidad parcial del sistema, el afectado no dispone de condiciones plenas para comprender por qué se produjo el resultado ni cómo cada variable incidió en su score. Segundo, por el vínculo

directo entre ese resultado y una solicitud de privación preventiva de la libertad. En ese contexto, la probabilidad deja de ser una mera ayuda descriptiva y puede convertirse en una forma de cosificación administrativa en donde el sujeto ya no es solo alguien respecto de quien se investiga un hecho, sino alguien cuyo futuro es anticipado y gestionado por una arquitectura estadística (Díaz et al,. 2024; Torres et al,. 2022).

Con todo, aquí subsisten algunas contenciones parciales, el sistema no decide formalmente, el resultado se presenta como insumo y la decisión final sigue radicada en el juez de control de garantías. Esas condiciones no eliminan la lesión potencial, pero sí impiden afirmar una cosificación absoluta.

8.4.4. Control humano significativo

En PRiSMA, el control humano existe formalmente en varios momentos. En el diseño, la herramienta es desarrollada y ajustada por equipos técnicos de la Fiscalía. En el uso operativo, el fiscal decide si genera y utiliza el reporte como insumo en la audiencia. Y en la decisión final, la medida de aseguramiento sigue dependiendo del juez de control de garantías conforme al artículo 308 de la Ley 906 de 2004. Desde esa perspectiva, PRiSMA no configura una automatización total de la decisión, pues el algoritmo no impone por sí mismo la detención preventiva ni desplaza formalmente al decisor judicial (Torres et al,. 2022).

Sin embargo, la suficiencia material de ese control humano es mucho más débil. A partir de la reconstrucción del caso, que no es posible verificar públicamente protocolos claros de apartamiento del puntaje, reglas internas sobre el peso del puntaje en la solicitud

fiscal, mecanismos sistemáticos de trazabilidad que permitan saber cuánto influyó el reporte en las decisiones o auditorías periódicas sobre sesgos y errores (Díaz et al,. 2024).

A ello se añade la opacidad parcial del modelo: si el fiscal o el juez no pueden comprender cómo se generó el resultado ni qué peso específico tuvieron las variables, el control humano corre el riesgo de degradarse a una supervisión formal sobre un output ya producido, sin capacidad real de control sobre la inferencia. En términos prácticos, el sesgo de automatización se vuelve plausible: el puntaje puede funcionar como evidencia técnica de fácil circulación, anclando la solicitud fiscal y modulando indirectamente la valoración judicial.

Por ello, aunque el sistema conserva un decisor humano en la etapa final, no puede afirmarse que el control sea plenamente significativo. Lo que existe es una intervención humana suficiente para evitar hablar de automatización total, pero insuficiente para garantizar una supervisión sustantiva, crítica y documentada del funcionamiento y peso real del puntaje.

8.4.5. Conclusión evaluativa del caso PRiSMA

Aplicados los umbrales reconstruidos en el Capítulo III a la evidencia técnico-institucional sistematizada en el Capítulo II, el caso PRiSMA presenta una tensión garantista alta. El sistema no cumple el estándar de no discriminación, debido al peso que adquieren proxies penales y territoriales en un esquema de predicción cuyo control externo es insuficiente. A su vez, cumple parcialmente los estándares de igualdad material, dignidad humana y control humano significativo: si bien existen elementos de contención formal —uso como insumo y decisión final en cabeza del juez—, esas salvaguardas resultan

insuficientes frente al carácter opaco del modelo, al uso de datos históricamente sesgados y a su inserción en decisiones de privación preventiva de libertad.

La lección constitucional del caso es especialmente relevante para el contexto colombiano. PRiSMA muestra que una herramienta de predicción penal puede presentarse como solución técnica a problemas de congestión y uso de cupos carcelarios, pero al mismo tiempo reconfigurar materialmente las garantías del Estado Social de Derecho si convierte trayectorias penales previas en fundamento probabilístico de la restricción de libertad, sin transparencia robusta, sin trazabilidad suficiente y sin controles humanos plenamente significativos.

En esa medida, el caso confirma la hipótesis central del trabajo: en contextos de poder punitivo históricamente selectivo, la IA estatal no puede evaluarse solo por su promesa de eficiencia, sino por los mecanismos concretos mediante los cuales reproduce o mitiga desigualdades y por el grado en que preserva o erosiona la igualdad material, la no discriminación y la dignidad humana.

8.5 Caso Fiscal Watson

En el caso de Fiscal Watson, la base fáctica relevante ya fue establecida en el Capítulo II: se trata de una herramienta de analítica de contenidos y búsqueda avanzada implementada por la Fiscalía General de la Nación, basada en tecnología IBM Watson, orientada a asociar noticias criminales y a producir insumos de gestión y priorización dentro del SPOA. También quedó demostrado que opera sobre grandes volúmenes de datos estructurados y no estructurados, que su salida consiste en asociaciones e hipótesis de conexión entre casos, que

su gobernanza se encuentra fragmentada entre proveedor tecnológico y entidad operadora, y que existen déficits importantes de transparencia, trazabilidad y evaluación integral de impacto, aun cuando se reporten resultados agregados de uso institucional. Sobre esa base, lo que corresponde ahora es determinar si esa configuración satisface o compromete las garantías sustantivas analizadas.

8.5.1. Igualdad material

El estándar aplicable exige que un sistema algorítmico estatal no reproduzca ni profundice desigualdades estructurales mediante la distribución asimétrica de cargas de vigilancia, sospecha o priorización investigativa sobre grupos o territorios históricamente vulnerados. La igualdad material no se satisface cuando el Estado toma como neutros datos y patrones que, en realidad, condensan trayectorias históricas de selectividad penal y social.

En Fiscal Watson, este umbral se ve comprometido porque la herramienta se alimenta del SPOA y de variables asociadas a georreferenciación del delito, modus operandi, reincidencia y relaciones entre casos. Tales insumos no provienen de una realidad social "pura", sino de un sistema de persecución penal que puede arrastrar sesgos históricos de vigilancia y control. Cuando esos datos se indexan, correlacionan y reutilizan para sugerir conexiones o priorizar investigaciones, existe el riesgo de que el sistema no solo refleje desigualdades previas, sino que las automatice y les otorgue una apariencia de objetividad técnica (Morales et al., 2021). La evidencia disponible ya advertía, en tu propia Fase III, que esta lógica puede traducirse en una automatización del sesgo histórico, especialmente cuando variables territoriales o institucionales funcionan como proxies de exclusión.

La tensión con la igualdad material se intensifica porque el sistema no se limita a recuperar información: su salida puede orientar qué casos se conectan, cuáles reciben prioridad y dónde se concentran recursos investigativos. En un contexto como el colombiano, donde el poder punitivo ha operado históricamente de manera selectiva sobre ciertos territorios y poblaciones, una herramienta que asocia masivamente registros y jerarquiza relevancias puede contribuir a reforzar esa selectividad si no cuenta con correctivos materiales suficientes (Palacios et al., 2024). Dado que la documentación pública no permite verificar evaluaciones ex ante robustas sobre igualdad ni mecanismos de mitigación suficientemente trazables, no es posible sostener que la herramienta neutralice tales riesgos de modo adecuado.

8.5.2. No discriminación

El estándar de no discriminación exige prevenir tanto discriminación directa como discriminación indirecta derivada del uso de proxies, datos sesgados o configuraciones que produzcan efectos desproporcionados sobre grupos protegidos. En sistemas de analítica penal, este estándar se activa cuando variables aparentemente neutrales —territorio, historial institucional, redes de relación, reiteración delictiva— pueden funcionar como sustitutos de condición socioeconómica, pertenencia territorial o selectividad policial.

En Fiscal Watson, la afectación a esta garantía se configura en ese plano indirecto. La propia caracterización técnico-institucional reconoce que variables operativas como georreferenciación, historial institucional y redes de interacción pueden funcionar como proxies de condiciones socioeconómicas o de selectividad policial, de manera que el sesgo no depende de introducir una categoría sensible explícita, sino de reutilizar correlaciones

estructurales incrustadas en el dato administrativo penal. En este punto, la herramienta no aparece como un sistema abiertamente discriminatorio, pero sí como una arquitectura capaz de trasladar y reorganizar desigualdades históricas en el proceso de asociación y priorización de casos (Flórez Rojas & Vargas Leal, 2022; Morales et al., 2021).

El problema se agrava por la ausencia de información pública suficiente sobre métricas de error, sesgo por subpoblaciones o auditorías externas independientes. La evidencia disponible permite conocer la finalidad general del sistema y algunos resultados agregados, pero no permite verificar de manera robusta si determinadas poblaciones o territorios quedan sobreexpuestos a asociaciones, alertas de relevancia o hipótesis de conexión. En términos garantistas, esta falta de verificabilidad impide descartar razonablemente la existencia de impactos diferenciados injustificados. Así, el riesgo de discriminación indirecta no puede considerarse neutralizado solo porque el sistema no use categorías sensibles explícitas; por el contrario, la ausencia de controles suficientes obliga a tratar este riesgo como estructural.

8.5.3. Dignidad humana

La dignidad humana exige que la persona no sea reducida a una etiqueta de riesgo, asociación estadística o hipótesis automatizada de relevancia estatal. En el contexto de analítica penal, este estándar impide que la singularidad del sujeto y la complejidad del caso sean reemplazadas por una racionalidad de eficiencia que instrumentaliza al individuo como objeto de gestión informacional.

En Fiscal Watson, la tensión con la dignidad se manifiesta en la propia lógica del sistema: la herramienta indexa, correlaciona y recupera información sobre personas

vinculadas a noticias criminales para sugerir relaciones entre casos y orientar hipótesis investigativas. Desde la perspectiva garantista, esto implica que el sujeto puede dejar de ser visto prioritariamente como persona singular y pasar a ser tratado como nodo de una red de datos, susceptible de asociación, priorización o conexión algorítmica. La afectación no depende de que el sistema "declare culpable" a alguien, sino de que convierta su trayectoria, sus relaciones o sus metadatos en materia prima para una gestión investigativa intensiva sin suficiente explicabilidad (Morales et al., 2021; Flórez Rojas & Vargas Leal, 2022). En la Fase III ya se había advertido con claridad que esta dinámica podía traducirse en una forma de instrumentalización del individuo en favor de una métrica de eficiencia administrativa.

La afectación a la dignidad se agrava por la opacidad. Al tratarse de una herramienta desarrollada por un tercero privado, protegida por reserva técnica y con documentación pública limitada sobre sus criterios de asociación, el sujeto afectado —y aun los propios operadores— no puede reconstruir plenamente por qué se generó una determinada asociación entre casos ni qué lógica de relevancia produjo el resultado. En ese contexto, la persona no solo es convertida en dato procesable, sino que además se ve privada de condiciones suficientes para comprender y controvertir racionalmente la premisa técnica que la vincula con determinadas hipótesis investigativas (Morales et al., 2021; Flórez Rojas & Vargas Leal, 2022). Esa combinación de cosificación y opacidad es incompatible con una comprensión fuerte de la dignidad humana en el Estado Social de Derecho.

Con todo, a diferencia de sistemas directamente orientados a puntuar libertad o a generar alertas de riesgo individualizadas, aquí existen algunas contenciones parciales: la salida de Fiscal Watson consiste en hipótesis de asociación que requieren verificación humana posterior, y no en conclusiones probatorias autosuficientes. Esa condición atenúa la

intensidad de la afectación, pero no la elimina, porque la reducción del sujeto a una asociación estadística puede incidir de manera material en la orientación de la investigación y en la relevancia institucional otorgada a su caso.

8.5.4. Control humano significativo

En esta investigación, el control humano significativo se asume como condición instrumental para hacer exigibles y controlables las garantías sustantivas. No basta con que un funcionario intervenga al inicio o al final de la cadena; se requiere una capacidad real de comprender, cuestionar, apartarse del output y dejar trazabilidad suficiente del uso del sistema. El estándar exige, por tanto, un control humano material y no meramente decorativo.

En Fiscal Watson, el control humano existe formalmente en varias fases. La formulación de consultas, la interpretación de resultados y la verificación posterior están a cargo de analistas e investigadores humanos; además, la salida del sistema no automatiza por sí misma decisiones jurídicas finales, sino que se orienta a apoyar hipótesis investigativas y priorizaciones operativas. También existe, al menos en teoría, un control ex post en la medida en que los equipos pueden revisar resultados, ajustar parámetros y corregir registros del SPOA (Flórez Rojas & Vargas Leal, 2022). En ese sentido, no se trata de una herramienta de automatización total.

Sin embargo, la suficiencia material del control humano es bastante más débil. La Fase III ya había señalado el riesgo de automatización de facto mediante un efecto de anclaje: el investigador puede tomar el reporte de Watson como una “verdad científica de base” y ajustar su juicio a partir de ella, otorgándole un peso injustificado solo por su origen tecnológico. Ese riesgo se ve reforzado por dos elementos adicionales: la opacidad del

funcionamiento interno y la ausencia de métricas públicas de desempeño, error o impacto en derechos. Si el operador humano no puede comprender plenamente cómo se produjo una asociación, ni cuenta con trazabilidad suficiente para revisar críticamente el proceso, su intervención se vuelve formalmente existente pero materialmente limitada.

En consecuencia, el control humano en Fiscal Watson no es inexistente, pero sí condicionado por asimetrías de información, por dependencia del proveedor y por falta de auditoría externa suficiente. Ello impide sostener que el estándar de CHS quede satisfecho de manera robusta. Lo que existe es una intervención humana real, pero debilitada por la falta de explicabilidad y por el riesgo de que la lógica analítica se naturalice como criterio predominante de priorización.

8.5.5. Conclusión evaluativa del caso Fiscal Watson

Aplicados los umbrales reconstruidos en el Capítulo III a la evidencia técnico-institucional sistematizada en el Capítulo II, el caso Fiscal Watson presenta una tensión garantista alta, aunque distinta de la observada en herramientas directamente orientadas a puntuar riesgo individual o a recomendar privación de libertad. El sistema no cumple el estándar de no discriminación, debido al uso de proxies penales y territoriales sin auditorías públicas suficientes sobre impacto diferencial; y cumple parcialmente los estándares de igualdad material, dignidad humana y control humano significativo, pues existen elementos de contención relevantes —interpretación y verificación humana, carácter instrumental del output, ausencia de automatización jurídica total—, pero estos no neutralizan los efectos de una infraestructura opaca de asociación y priorización fundada en datos históricamente sesgados.

La lección constitucional del caso es especialmente importante para el contexto colombiano. Fiscal Watson muestra que, aun cuando una herramienta no "decida" jurídicamente por sí sola, puede reconfigurar de manera intensa el poder investigativo del Estado si indexa masivamente relatos y metadatos, opera con componentes propietarios poco auditables y orienta la priorización institucional a partir de patrones que pueden reproducir selectividad histórica. En esa medida, el caso confirma la hipótesis central de esta investigación: la validez constitucional de la IA estatal no depende solo de su promesa de eficiencia o de sus resultados agregados, sino de los mecanismos concretos mediante los cuales organiza datos, distribuye atención investigativa, produce asociaciones relevantes y conserva —o debilita— el control humano, la igualdad material, la no discriminación y la dignidad humana.

8.6 Resultado de la evaluación garantista por caso

8.6.1. Criterios de valoración del juicio garantista

La valoración que se presenta a continuación no se apoya en una lógica puramente cuantitativa ni en una escala numérica rígida, pues ello sería inconsistente con el propio objeto de estudio, en donde se analizaron sistemas complejos, opacos y heterogéneos, cuyas afectaciones no siempre son plenamente medibles mediante indicadores numericos uniformes. En su lugar, se adopta una escala cualitativa de tres niveles: cumple (C), cumple parcialmente (CP) y no cumple (NC), lo que permite traducir los hallazgos técnicos, institucionales y normativos en una conclusión garantista comprensible, trazable y comparable entre casos.

En el mismo sentido, se hace la aclaración que el juicio emitido en esta punto es un juicio de compatibilidad garantista, no un juicio de eficiencia técnica, utilidad operativa o conveniencia administrativa. En otras palabras, no se evalúa si el sistema "funciona bien" desde la lógica institucional que lo promueve, sino si su funcionamiento, tal como ha sido reconstruido en los capítulos anteriores, resulta compatible con las exigencias de igualdad material, no discriminación, dignidad humana y control humano significativo.

Por ello, un sistema puede presentarse como técnicamente útil o institucionalmente eficiente y, aun así, ser valorado como incompatible o parcialmente incompatible con las garantías del Estado Social de Derecho.

Ahora bien, sobre el sentido de cada uno de los niveles o categorías de escala, se debe tener en cuenta lo siguiente:

La categoría cumple (C) se utiliza cuando la evidencia disponible permite concluir que el sistema, en la garantía analizada, satisface de manera suficiente el umbral mínimo exigible, ya sea porque:

- el diseño y el despliegue muestran salvaguardas robustas y verificables;
- existen mecanismos adecuados de control, trazabilidad o corrección;
- no se advierten desviaciones estructurales relevantes;
- o, aun existiendo tensiones menores, estas no alcanzan a comprometer de forma significativa el núcleo de la garantía.

No se exige perfección absoluta para calificar un caso como "cumple". Lo determinante es que, a la luz de la evidencia disponible, las salvaguardas sean materialmente suficientes para evitar que la garantía se vea erosionada en su aplicación real.

En cuanto a si la categoría cumple parcialmente (CP) se utiliza cuando el caso muestra una compatibilidad incompleta o inestable con el umbral aplicable. Esto ocurre, por ejemplo, cuando:

- existen salvaguardas formales, pero su eficacia es limitada;
- hay intervención humana, pero esta no alcanza a ser plenamente significativa;
- se observan mecanismos de contención, aunque no neutralizan completamente la afectación;
- la configuración presenta elementos de cumplimiento y, al mismo tiempo, desviaciones relevantes que impiden afirmar una conformidad plena.

Esta categoría es especialmente importante porque reconoce una zona intermedia entre la plena compatibilidad y la incompatibilidad abierta. Su función es evitar dos extremos igualmente problemáticos: por un lado, considerar que basta una salvaguarda formal para concluir que el sistema "cumple"; por otro, tratar toda tensión o insuficiencia como si equivaliera automáticamente a un incumplimiento absoluto.

En términos metodológicos, la categoría "cumple parcialmente" indica que el sistema no es enteramente incompatible, pero tampoco ofrece una satisfacción robusta del estándar. Existe una adecuación parcial, condicionada o precaria, que debe ser explicitada mediante la identificación de la desviación principal.

Y por último, la categoría no cumple (NC) se utiliza cuando la evidencia muestra una incompatibilidad sustancial o estructural entre la configuración del sistema y el umbral exigible de la garantía. Ello ocurre cuando:

- la afectación es directa o intensa;

- la ausencia de salvaguardas es grave;
- la opacidad impide control efectivo;
- existen impactos o riesgos suficientemente documentados que comprometen el núcleo de la garantía;
- la propia estructura del sistema hace inviable considerar satisfecho el umbral, aun si subsisten ciertos correctivos menores.

Esta categoría no exige necesariamente la existencia de una sentencia que declare ilícito el sistema, aunque la existencia de decisiones judiciales de suspensión o anulación constituye una evidencia muy fuerte. También puede utilizarse cuando, aun sin pronunciamiento judicial definitivo, el conjunto de hallazgos empíricos permite concluir que la herramienta opera en tensión estructural con las exigencias mínimas del estándar constitucional.

De otro lado, la asignación del juicio no se hizo de manera arbitraria, sino a partir de cinco criterios de valoración, aplicados en cada garantía según su naturaleza:

i) Intensidad de la afectación

Se valoró qué tan directamente la configuración del sistema compromete la garantía. No es lo mismo un riesgo remoto o hipotético que una afectación estructural, reiterada o ya documentada. A mayor intensidad de la afectación, mayor probabilidad de un juicio de “no cumple”.

ii) Grado de verificabilidad pública

Se consideró si el sistema ofrece condiciones suficientes de transparencia, trazabilidad y acceso a la información relevante para permitir control institucional, judicial

o ciudadano. Cuando elementos esenciales del funcionamiento permanecen opacos o no verificables, ello se valoró negativamente, no como un vacío neutral, sino como un hallazgo de gobernanza que debilita el estándar.

iii) Existencia y robustez de salvaguardas

Se examinó si existían mecanismos reales —y no meramente nominales— de mitigación, auditoría, control humano, revisión, contradicción o corrección. La presencia de salvaguardas formales podía mejorar la valoración, pero solo en la medida en que fueran materialmente eficaces.

iv) Presencia de impactos o riesgos documentados

Se tuvo en cuenta la existencia de evidencia empírica, técnica, doctrinal o judicial sobre falsos positivos, disparidades, estigmatización, afectaciones en subgrupos, déficits de transparencia o vulneraciones de derechos. La fuerza del juicio aumenta cuando esos impactos están documentados de manera consistente.

v) Capacidad de contradicción y control humano

En especial para dignidad, debido proceso y CHS, se valoró si el operador humano y el sujeto afectado podían realmente comprender, cuestionar y eventualmente apartarse del resultado producido por el sistema. Cuando la intervención humana era meramente formal, o cuando la lógica algorítmica no podía ser comprendida ni discutida, ello incidió negativamente en la valoración.

Seguidamente, el bloque “desviación principal” cumple la función de sintetizar cuál es el mecanismo concreto que explica el resultado, es decir, representa la razón central por

la cual el sistema cumple solo parcialmente o no cumple. Junto a la desviación principal, se encuentra el cuadro denominado efecto sobre la garantía. Este elemento responde sintetiza la cuestión de ¿qué produce concretamente esa desviación en la garantía evaluada? La respuesta permite traducir la arquitectura técnico-institucional en una afectación jurídico-constitucional concreta.

Finalmente, debe precisarse que la asignación de un juicio de cumplimiento no supone declarar de manera abstracta la constitucionalidad o inconstitucionalidad total del sistema examinado, ni sustituye un control judicial formal. Se trata, más bien, de una valoración académica y argumentada de compatibilidad garantista, formulada a partir de la evidencia pública y de los estándares reconstruidos en la investigación. Por eso, cuando la información disponible no permitía verificar ciertos elementos esenciales del sistema, esa falta de verificabilidad fue tratada como un hallazgo relevante y no como una licencia para presumir legitimidad.

8.6.2. Tablas de síntesis de la evaluación garantista

8.6.2.1 Igualdad material

Escala de juicio: C = cumple | CP = cumple parcialmente | NC = no cumple.

Tabla 10
Evaluación garantista de los casos: igualdad material

Caso	Evidencia relevante	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
SyRI	Focalización territorial en barrios empobrecidos; cruce masivo de datos de bienestar; exposición reforzada de residentes y	NC	La pobreza y el territorio operan como criterios de sobreexposición al control administrativo	Automatiza desigualdades estructurales y convierte la vulnerabilidad social en

Caso	Evidencia relevante	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
	beneficiarios en zonas vulnerables.		sin correctivos materiales suficientes.	umbral de sospecha.
AFR Locate	Selección policial de lugares de despliegue y watchlists; vigilancia biométrica de transeúntes en espacios públicos; discrecionalidad relevante sobre quién y dónde vigilar.	NC	La elección de zonas y objetivos puede concentrar vigilancia sobre espacios y grupos ya sometidos a mayor escrutinio policial.	Redistribuye de manera desigual la carga de vigilancia en el espacio público.
COMPAS	Uso en decisiones de supervisión, clasificación, manejo de casos e informes pre-sentenciales; datos penales con huella histórica; evidencia pública de cargas asimétricas sobre población hipervigilada.	CP	El sistema racionaliza restricciones sobre sujetos ya expuestos a selectividad penal sin correctivos materiales plenamente verificables.	Refuerza desigualdades estructurales mediante una apariencia de neutralidad técnica.
PRiSMA	Predicción de reincidencia para sustentar medidas de aseguramiento; uso de registros penales e institucionales históricos; despliegue verificable en entornos urbanos de alta intensidad punitiva.	CP	Convierte trayectorias producidas por un sistema penal desigual en fundamento probabilístico de restricción preventiva de libertad.	Riesgo de profundizar desigualdades sociales y territoriales bajo el discurso de eficiencia.
Fiscal Watson	Asociación y priorización de casos a partir del SPOA, georreferenciación y variables institucionales; orientación de recursos investigativos según patrones previos.	CP	Reutiliza datos penales históricamente selectivos para jerarquizar atención investigativa sin correctivos	Puede reforzar desigualdades territoriales y sociales en la distribución de la persecución penal.

Caso	Evidencia relevante	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
			materiales robustos.	

8.6.2.2 2. No discriminación

Escala de juicio: C = cumple | CP = cumple parcialmente | NC = no cumple.

Tabla 11
Evaluación garantista de los casos: no discriminación

Caso	Evidencia relevante (Tabla 1)	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
SyRI	Uso de datos administrativos socioeconómicos y focalización territorial en zonas vulnerables; ausencia de mitigación pública suficiente del riesgo discriminatorio.	NC	Variables aparentemente neutrales operan como proxies de pobreza, origen migratorio y territorialidad.	Reproduce discriminación indirecta bajo apariencia de eficiencia administrativa.
AFR Locate	Posibles sesgos por raza y sexo; watchlists y despliegue territorial; incumplimiento del Deber de Igualdad en el Sector Público en la verificación independiente del impacto diferencial.	NC	No se acreditó una evaluación independiente suficiente sobre sesgos ni una mitigación robusta del impacto diferencial.	Expone desproporcionadamente a mujeres y minorías étnicas a errores y vigilancia selectiva.
COMPAS	Variables correlacionadas con raza y estatus socioeconómico (historial, empleo, educación, residencia); debate	NC	Los proxies estructurales trasladan sesgos del sistema penal al score sin mitigación	Normaliza la discriminación indirecta dentro de decisiones penales de alto impacto.

Caso	Evidencia relevante (Tabla 1)	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
	documentado sobre falsos positivos y negativos asimétricos.		pública plenamente controlable.	
PRiSMA	Registros de Policía, Fiscalía e INPEC; proxies penales y territoriales; falta de auditorías externas completas sobre sesgo e impacto diferencial.	NC	La herramienta puede proyectar sesgos institucionales históricos sin un esquema verificable de mitigación.	Riesgo de discriminación indirecta en decisiones de detención preventiva.
Fiscal Watson	Georreferenciación, historial institucional, redes y patrones de asociación; ausencia de auditorías públicas suficientes sobre impacto diferencial por subgrupos.	NC	Los proxies penales y territoriales pueden reorganizar selectividad histórica en la priorización investigativa.	Reproduce patrones desiguales de vigilancia y persecución penal bajo neutralidad analítica.

8.6.2.3 3. Dignidad humana

Escala de juicio: C = cumple | CP = cumple parcialmente | NC = no cumple.

Tabla 12
Evaluación garantista de los casos: dignidad humana

Caso	Evidencia relevante (Tabla 1)	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
SyRI	Generación de alertas de riesgo opacas sobre personas y hogares; despliegue en barrios estigmatizados; baja	NC	Reduce al sujeto a una alerta probabilística dentro de una cartografía	Cosificación y estigmatización del ciudadano como objeto de gestión

Caso	Evidencia relevante (Tabla 1)	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
	explicabilidad para el afectado.		territorial de sospecha.	administrativa del riesgo.
AFR Locate	Captura biométrica masiva de transeúntes y producción de posibles matches; reglas de borrado para no coincidencias y verificación humana posterior.	CP	El cuerpo se transforma en dato procesable de vigilancia, aunque existan contenciones formales parciales.	Debilita el anonimato y la autonomía en el espacio público, sin llegar a una cosificación absoluta en todos los casos.
COMPAS	Salida en deciles y categorías de riesgo; integración del puntaje en decisiones restrictivas; advertencias judiciales sobre su alcance grupal.	CP	La persona es reducida a una etiqueta actuarial que puede desplazar la valoración individualizada.	Riesgo de cosificación y estigmatización mediante el puntaje de riesgo.
PRiSMA	Reporte probabilístico de riesgo de reincidencia para apoyar solicitud de medida de aseguramiento antes de condena; decisión final aún radica en el juez.	CP	El imputado es leído como perfil de riesgo antes de condena, aunque el resultado se presente como insumo y no como decisión autónoma.	Debilita la consideración del sujeto como persona singular y no solo como objeto de gestión del riesgo.
Fiscal Watson	Asociaciones e hipótesis de conexión entre casos y personas; opacidad de la lógica analítica; verificación posterior por investigadores.	CP	El sujeto se convierte en nodo de relevancia analítica dentro de una arquitectura opaca de asociación.	Instrumentaliza a la persona en función de la eficiencia investigativa, aunque sin automatización jurídica total.

8.6.2.4 Control humano significativo (CHS)

Escala de juicio: C = cumple | CP = cumple parcialmente | NC = no cumple.

Tabla 13
Evaluación garantista de los casos: control humano significativo

Caso	Evidencia relevante (Tabla 1)	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
SyRI	Intervención humana formal en solicitud, operación y retroalimentación; opacidad del modelo; acceso restringido para el afectado y baja capacidad de contradicción.	NC	La intervención humana existe, pero no puede comprender ni controlar materialmente la lógica de marcación del sistema.	El control humano se degrada a una función decorativa dentro de una arquitectura opaca.
AFR Locate	Autorización previa, revisión humana de cada coincidencia y control judicial posterior; algoritmo propietario y riesgo de sesgo de automatización.	CP	La verificación humana existe, pero opera sobre un resultado opaco que puede anclar la intervención policial.	El control es real pero insuficiente para neutralizar plenamente el peso de la alerta algorítmica.
COMPAS	Uso formal como insumo; posibilidad de apartamiento; advertencias judiciales de no determinación; opacidad y anclaje práctico del puntaje.	CP	El decisor humano conserva la última palabra, pero la caja negra y el sesgo de automatización vacían de contenido parte del control.	Riesgo de automatización de facto y de supervisión meramente formal.
PRiSMA	Fiscal usa el reporte y el juez decide la medida; no hay protocolos plenamente verificables de apartamiento ni trazabilidad suficiente del peso real del score.	CP	La presencia de un decisor humano no elimina la dependencia práctica frente a un resultado	Debilita el control humano material y favorece una delegación

Caso	Evidencia relevante (Tabla 1)	Juicio (C/CP/NC)	Desviación principal	Efecto sobre garantía
			difícilmente explicable y controlable.	encubierta del juicio.
Fiscal Watson	Analistas e investigadores formulan consultas, interpretan resultados y verifican asociaciones; opacidad propietaria y ausencia de métricas públicas robustas.	CP	La intervención humana existe, pero está condicionada por la dependencia epistémica respecto del proveedor y por la falta de trazabilidad suficiente.	Puede convertir al operador en validador de una salida analítica difícil de cuestionar críticamente.

9. Conclusiones

La investigación permitió establecer que la incorporación de sistemas de inteligencia artificial en dispositivos estatales de poder, vigilancia y control no reconfigura las garantías del Estado Social de Derecho de manera abstracta ni uniforme, sino a través de configuraciones sociotécnicas concretas. El estudio demuestra que la afectación o preservación relativa de la igualdad material, la no discriminación y la dignidad humana depende menos de la mera existencia de un sistema de IA y más de la combinación específica entre tipo de datos, arquitectura del modelo, reglas de uso en la decisión estatal, grado de opacidad, arreglos de gobernanza y densidad del control humano significativo. Dicho de otro modo: la IA estatal no es un "factor externo" al Estado Social de Derecho, sino una forma contemporánea de reorganizar sus prácticas de vigilancia, selección, priorización y castigo; por ello, su análisis jurídico exige examinar cómo está diseñada y qué efectos distribuye, no solo para qué dice haber sido creada. Esta conclusión responde de manera directa a la

pregunta de investigación y confirma la hipótesis general del trabajo: las promesas de eficiencia y neutralidad técnica tienden a ocultar mecanismos de reproducción de desigualdad cuando no están acompañadas de salvaguardas robustas, verificables y orientadas a derechos.

Desde la perspectiva de los objetivos específicos, el primero se cumplió en la medida en que la investigación reconstruyó un marco conceptual capaz de leer la IA estatal no como herramienta neutral, sino como dispositivo de poder, vigilancia y control. El segundo se satisfizo mediante la caracterización homogénea de cinco casos —COMPAS, SyRI, AFR Locate, PRiSMA y Fiscal Watson— usando una misma matriz técnico-institucional que permitió mapear finalidades, datos, modelos, salidas, roles decisionales, gobernanza, transparencia, control humano y evidencias de impacto. El tercero se alcanzó al analizar esos hallazgos para reconstruir umbrales mínimos de validez constitucional aplicables a igualdad material, no discriminación, dignidad humana y control humano significativo. Finalmente, el cuarto objetivo se cumplió al evaluar cada caso con base en esos umbrales, determinando niveles de cumplimiento, cumplimiento parcial o incumplimiento, y precisando las desviaciones relevantes en cada configuración. En ese sentido, el diseño metodológico sí logró "cerrar el ciclo" entre teoría, descripción empírica, análisis integral y evaluación garantista, tal como había sido propuesto desde la metodología inicial.

Uno de los hallazgos más consistentes del estudio es que ninguno de los casos analizados satisface plenamente el conjunto de garantías evaluadas. El patrón intercasos muestra, en primer lugar, que la no discriminación es la garantía más intensamente comprometida. En los cinco casos examinados aparecen, con distinta configuración, problemas asociados al uso de proxies, a la reutilización de datos institucionales con huellas

históricas de vigilancia o exclusión, o a la ausencia de auditorías públicas robustas sobre impacto diferencial. En segundo lugar, la igualdad material también resulta sistemáticamente tensionada, ya sea por focalización territorial (SyRI, AFR Locate), por racionalización actuarial de cargas penales (COMPAS, PRiSMA) o por priorización investigativa basada en correlaciones históricas (Fiscal Watson), el efecto recurrente es la posibilidad de redistribuir vigilancia, sospecha o restricción de libertad sobre poblaciones previamente expuestas al control estatal. En tercer lugar, la dignidad humana se erosiona cuando las personas son reducidas a puntajes, alertas, plantillas biométricas o asociaciones estadísticas que desplazan la valoración individualizada. Finalmente, el control humano significativo se muestra, en la mayoría de los casos, como una garantía formalmente invocada pero materialmente debilitada por opacidad, anclaje cognitivo y dependencia institucional respecto del output. Esta cartografía confirma que la reconfiguración garantista operada por la IA estatal se produce, sobre todo, por la interacción entre opacidad, datos históricamente sesgados, discrecionalidad institucional y debilidad de los controles.

La comparación de casos permitió identificar un conjunto de mecanismos recurrentes de reproducción de desigualdades. El primero es la opacidad estructural, es decir, cuando no son verificables públicamente la lógica del modelo, sus ponderaciones, sus umbrales o sus reglas de correlación, la ciudadanía, los afectados y aun los propios operadores jurídicos quedan impedidos para controlar de manera material el output. El segundo es el uso de proxies sensibles, incluso sin introducir categorías como raza, pobreza o migración de manera explícita, el sistema puede operar con variables funcionalmente equivalentes que trasladan desigualdades históricas al presente decisional. El tercero es la discrecionalidad institucional sin criterios verificables, visible en la selección de watchlists y zonas de

despliegue en AFR Locate, en la focalización territorial de SyRI, en la incorporación práctica del puntaje en COMPAS y PRiSMA, o en la asociación priorizada de noticias criminales en Fiscal Watson. El cuarto es la debilidad del control humano significativo, que en varios casos se presenta más como "firma humana" sobre resultados opacos que como supervisión real capaz de comprender, cuestionar y revertir la inferencia automatizada. Y el quinto es la ausencia o insuficiencia de recursos efectivos de contradicción, especialmente cuando el afectado no sabe por qué fue marcado, priorizado, asociado o perfilado. Estos mecanismos, que el estudio había anticipado como relevantes desde el planteamiento del problema, aparecieron de manera transversal y confirmaron que la erosión de garantías no es un accidente marginal, sino una posibilidad estructural del uso estatal de IA en contextos de poder, vigilancia y castigo.

Al mismo tiempo, la investigación permitió identificar algunos mecanismos de contención o mitigación, aunque ninguno de ellos resultó suficiente por sí mismo para neutralizar completamente los riesgos. Entre estos mecanismos aparecen: la existencia de evaluaciones de impacto o documentos de gobernanza; las advertencias judiciales sobre límites inferenciales y no determinación; las reglas de retención y borrado; ciertos márgenes de apartamiento o revisión humana; y algunas formas de trazabilidad documental. Sin embargo, el contraste comparado muestra que estos mecanismos solo contribuyen a preservar garantías cuando se presentan de manera conjunta, robusta y verificable.

Por el contrario, cuando actúan de forma aislada o reactiva por ejemplo, como respuesta judicial ex post o como documentación interna sin auditoría independiente, su capacidad de mitigación es limitada. El hallazgo, entonces, no es que la IA estatal sea inevitablemente incompatible con el Estado Social de Derecho, sino que su compatibilidad

depende de condiciones exigentes que en los casos analizados no aparecieron satisfechas de manera plena. Esto confirma la pertinencia del enfoque de la tesis, más que preguntar si la IA "sirve" o "no sirve", la cuestión jurídicamente relevante es bajo qué condiciones una práctica sociotécnica estatal puede ser considerada constitucionalmente aceptable.

En relación con la literatura previa, los hallazgos del estudio coinciden con buena parte de la doctrina y de los reportes que advertían que la IA estatal tiende a reproducir sesgos e inequidades cuando se nutre de datos históricos y opera bajo lógicas opacas. La investigación confirma, en línea con la literatura crítica ya revisada, que la promesa de neutralidad técnica suele encubrir decisiones previas sobre qué datos son relevantes, qué comportamientos se vuelven visibles y qué formas de riesgo merecen ser administradas. Sin embargo, el aporte propio del trabajo no se limita a reiterar esa crítica general, sino a traducirla en criterios jurídicos verificables. A diferencia de enfoques centrados exclusivamente en la ética algorítmica, la gobernanza de la innovación o la precisión técnica, esta tesis reorganiza el debate desde el Estado Social de Derecho y muestra que la validez de estas herramientas debe medirse por su incidencia concreta sobre igualdad material, no discriminación y dignidad humana, así como por la densidad del control humano significativo. En ese sentido, el estudio cubre un vacío identificado por el propio proyecto según el cual no bastaba con describir riesgos o denunciar sesgos; era necesario reconstruir umbrales constitucionales de control y aplicarlos comparativamente a casos concretos.

Las implicaciones teóricas del estudio son significativas. Primero, muestran que el análisis jurídico de la IA estatal no puede agotarse en la clásica oposición entre "decisión humana" y "decisión automatizada", porque muchos de los sistemas estudiados no sustituyen formalmente al humano, pero sí producen automatización de facto por anclaje, priorización

o dependencia institucional del resultado. Segundo, ponen en evidencia que la categoría de control humano significativo no debe leerse como una cláusula meramente procedimental, sino como una condición preventiva y correctiva para hacer exigibles las garantías sustantivas. Tercero, permiten reformular la discusión sobre igualdad y no discriminación más allá de la presencia explícita de variables sensibles, desplazando la atención hacia los proxies, la calidad institucional de los datos y la forma en que el sistema redistribuye cargas y beneficios. En cuarto lugar, el estudio sugiere que la dignidad humana resulta especialmente útil como criterio para examinar procesos de cosificación, estigmatización y administración probabilística del sujeto, allí donde el lenguaje de la eficiencia tiende a borrar la singularidad de la persona. Con ello, la tesis aporta una lectura constitucional más fina de la IA estatal, compatible con la crítica sociológica del control y del castigo, pero traducida en claves dogmáticas y operativas.

En el plano práctico y jurídico, la investigación no se propuso diseñar un catálogo de reformas ni un recetario de política pública, y debe mantenerse fiel a esa delimitación. Sin embargo, sí deja una conclusión práctica importante y es que en ausencia de transparencia operativa, trazabilidad, auditorías, evaluación de impacto y control humano verificable, la introducción de IA en funciones estatales de vigilancia, priorización penal o gestión de bienestar tiende a erosionar las garantías que el Estado Social de Derecho debe proteger.

En consecuencia, el aporte práctico del trabajo consiste en ofrecer una rejilla de lectura para que operadores jurídicos, jueces, fiscales, defensores, organismos de control y academia puedan identificar de manera más precisa cuándo una herramienta tecnológica opera como mitigación razonable y cuándo, por el contrario, funciona como dispositivo de dominación tecnificada. Esto se ajusta, además, a la justificación original del estudio, que no

buscaba prescribir soluciones regulatorias cerradas, sino aportar claridad analítica y criterios jurídicos verificables para valorar la conformidad o disconformidad de estas prácticas con el orden constitucional.

La investigación también deja ver algunos resultados no completamente esperados. El primero es que la mayor fragilidad comparada no se concentró exclusivamente en la dignidad humana, como podría sugerirse desde una lectura más filosófica del problema, sino en la no discriminación y en la debilidad del control humano significativo. Es decir, los casos mostraron que la lesión más recurrente no se presenta solo porque la persona sea cosificada, sino porque el sistema reorganiza diferencialmente la sospecha, la vigilancia y la restricción usando proxies y datos históricamente sesgados, mientras el control humano se debilita por opacidad y asimetría informacional. El segundo hallazgo inesperado es que los casos no se ordenan simplemente por "más o menos automatizados": sistemas que formalmente solo producen alertas, asociaciones o prioridades —como Fiscal Watson o AFR Locate— pueden reconfigurar intensamente garantías sin decidir jurídicamente por sí mismos. El tercero es que la comparación por configuraciones sociotécnicas resultó más fructífera que una comparación clásica entre países o marcos legales, porque permitió identificar patrones transversales de erosión y de mitigación que no dependen exclusivamente de la jurisdicción. Esa constatación valida el ajuste metodológico que fuiste consolidando durante el proceso y muestra que la tesis ganó consistencia al comparar no tanto "Estados" como ensamblajes de datos, modelos, reglas de uso y gobernanza.

Como toda investigación, este estudio presenta limitaciones que deben reconocerse con honestidad. La primera deriva de la propia naturaleza documental y comparada del diseño, en varios casos, especialmente en los nacionales y en los sistemas propietarios, la

información pública disponible es incompleta, fragmentaria o de difícil verificación, lo que obligó a registrar como "no verificable públicamente" aspectos importantes del funcionamiento interno, de los logs, de la calibración o del uso operativo real. La segunda limitación es que el análisis se concentró en cinco casos seleccionados por relevancia constitucional y diversidad de configuración, lo cual resulta metodológicamente adecuado para una tesis de grado, pero no agota la variedad de usos estatales de IA posibles. La tercera es que, al priorizar un enfoque jurídico-garantista, el trabajo no desarrolla auditorías técnicas de desempeño, experimentos cuantitativos ni trabajo de campo con operadores, por lo que algunas inferencias sobre uso práctico deben permanecer ancladas a evidencia documental y no a observación directa. La cuarta es que el capítulo de conclusiones no pretende resolver definitivamente debates abiertos en teoría de equidad, explicabilidad o regulación global de la IA, sino ubicar esos debates dentro de una cartografía constitucional concreta. Reconocer estas limitaciones no debilita el estudio; al contrario, permite delimitar con mayor precisión el alcance de sus afirmaciones.

A partir de lo anterior, se abren varias proyecciones y líneas de investigación futuras. En primer lugar, sería valioso profundizar en estudios empíricos sobre el uso real de estas herramientas por parte de operadores estatales, con especial atención al peso efectivo del output en la decisión final y a los sesgos cognitivos asociados al control humano. En segundo lugar, se requiere investigar con mayor detalle la forma en que métricas técnicas de equidad pueden o no dialogar con estándares jurídicos de igualdad material y no discriminación, sin reducir unas a otras. En tercer lugar, futuras investigaciones podrían ampliar el universo de casos hacia otros campos del Estado —educación, salud, empleo público, política social, migración— para determinar si los patrones identificados aquí se replican fuera del ámbito

penal y de vigilancia. En cuarto lugar, sería pertinente desarrollar análisis específicos sobre debido proceso algorítmico y explicabilidad, especialmente en el contexto colombiano, donde la práctica institucional parece avanzar con mayor rapidez que la capacidad de escrutinio público. Por último, una línea especialmente prometedora consiste en profundizar la categoría de control humano significativo como derecho o faceta garantista autónoma, articulándola con el problema de la caja negra y con la exigencia de recursos efectivos. Estas proyecciones son consistentes con la justificación metodológica y teórica del trabajo, que concibe sus matrices y categorías no como solución cerrada, sino como esquema replicable para nuevas investigaciones.

En definitiva, la tesis permite concluir que la incorporación de sistemas de IA en dispositivos estatales de poder, vigilancia y control reconfigura las garantías del Estado Social de Derecho principalmente a través de cuatro mecanismos: datos con huellas históricas de desigualdad, modelos opacos o insuficientemente auditables, reglas de uso que desplazan la sospecha o la priorización hacia ciertos grupos y territorios, y control humano insuficiente para contradecir, corregir o revertir el output. Cuando estos elementos se combinan, la tecnología no mitiga desigualdades: las reorganiza, las automatiza o las vuelve más difíciles de oponer. Por el contrario, solo bajo condiciones robustas de transparencia, trazabilidad, auditoría, evaluación de impacto y control humano significativo podría hablarse de una incorporación compatible con el horizonte garantista del Estado Social de Derecho. Así, la principal conclusión del trabajo no es que la IA estatal sea en sí misma ilícita o deseable, sino que su validez constitucional depende de condiciones exigentes que los casos analizados, en distinto grado, no lograron satisfacer plenamente. Esa es, al final, la respuesta más precisa que la investigación ofrece a la pregunta que la originó.

BIBLIOGRAFÍA

9.1 Fuentes normativas y Jurisprudenciales

Asamblea General de las Naciones Unidas. (2024). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development* (Resolución A/RES/78/265). Naciones Unidas. https://undocs.org/en/A/RES/78/265

Congreso de la República de Colombia. (2025). *Proyecto de Ley 043 de 2025 Senado: Por medio de la cual se regula la Inteligencia Artificial en Colombia para garantizar su desarrollo ético, responsable, competitivo e innovador, y se dictan otras disposiciones*.

Consejo de Europa. (2024). *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law* (CETS No. 225). Council of Europe. https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225

Consejo Superior de la Judicatura. (2024, 16 de diciembre). *Acuerdo PCSJA24-12243: Por el cual se adoptan lineamientos para el uso y aprovechamiento respetuoso, responsable, seguro y ético de la inteligencia artificial en la Rama Judicial*. Rama Judicial. https://www.ramajudicial.gov.co/documents/164295699/174277484/PCSJA24-12243.-USO-Y-APROVECHAMIENTO-DE-IA-EN-LA-RAMA-JUDICIAL.pdf

Corte Constitucional de Colombia. (2024, 2 de agosto). *Sentencia T-323 de 2024*. https://www.corteconstitucional.gov.co/relatoria/2024/T-323-24.htm

Corte Constitucional de Colombia. (2025). *Sentencia T-067 de 2025*. https://www.corteconstitucional.gov.co/relatoria/2025/T-067-25.htm

Corte Suprema de Justicia, Sala de Casación Civil, Agraria y Rural. (2025, 5 de noviembre). *Sentencia STC17832-2025* (Rad. 11001-02-03-000-2025-05001-00).

Departamento Nacional de Planeación. (2019). *Documento CONPES 3975: Política nacional para la transformación digital e inteligencia artificial*. DNP. https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3975.pdf

Departamento Nacional de Planeación. (2025). *Documento CONPES 4144: Política nacional de inteligencia artificial*. DNP. https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/4144.pdf

G7. (2023a). *Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems*. Author. https://www.g7hiroshimaprocess.org/docs/international-guiding-principles

G7. (2023b). *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*. Author. https://www.g7hiroshimaprocess.org/docs/international-code-of-conduct

International Organization for Standardization, & International Electrotechnical Commission. (2023a). *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system*. ISO. https://www.iso.org/standard/81230.html

International Organization for Standardization, & International Electrotechnical Commission. (2023b). *ISO/IEC 23894:2023 Information technology — Artificial intelligence — Risk management*. ISO. https://www.iso.org/standard/77304.html

Ministerio de Tecnologías de la Información y las Comunicaciones. (2022). *Guía con lineamientos generales para el uso de tecnologías emergentes*. MinTIC. https://gobiernodigital.mintic.gov.co/692/articles-179148_Guia_Tecnologias_Emergentes.pdf

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). NIST. https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). NIST. https://doi.org/10.6028/NIST.AI.600-1

Organización de Cooperación y Desarrollo Económicos. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). OCDE. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO]. (2021). *Recommendation on the Ethics of Artificial Intelligence*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Parlamento Europeo y Consejo de la Unión Europea. (2024). *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión*. Diario Oficial de la Unión Europea. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689

Procuraduría General de la Nación, & Defensoría del Pueblo. (2025, 30 de septiembre). *Directiva Conjunta No. 007: Estándares sobre transparencia algorítmica para los sistemas algorítmicos utilizados por el Estado*. Autor.

R (Bridges) v Chief Constable of South Wales Police & Ors [2020] EWCA Civ 1058 (Court of Appeal (Civil Division), 11 August 2020).

Rechtbank Den Haag. (2020). *ECLI:NL:RBDHA:2020:1878* (SyRI; versión en inglés).

State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

Superintendencia de Industria y Comercio. (2024, 21 de agosto). *Circular Externa No. 002 de 2024: Lineamientos sobre el tratamiento de datos personales en sistemas de inteligencia artificial*. SIC.

9.2 Fuentes doctrinales

Abásolo, E. (2023). Metodología de la investigación científica en derecho: Principios, criterios, técnicas. Dykinson.

AlgorithmWatch. (2020, 6 de abril). How Dutch activists got an invasive fraud detection algorithm banned. https://algorithmwatch.org/en/story/how-dutch-activists-got-an-invasive-fraud-detection-algorithm-banned/

Alston, P. (2019). Brief by the United Nations Special Rapporteur on extreme poverty and human rights as Amicus Curiae in the case of NJCM c.s./De Staat der Nederlanden

(SyRI) before the District Court of The Hague (case number: C/09/550982/ HA ZA 18/388). Naciones Unidas

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016a, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. Recuperado de: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016b, July 29). Technical response to Northpointe. ProPublica. Recuperado de: https://www.propublica.org/article/technical-response-to-northpointe

Aponte Fonseca, Y. Y. (2022). *Tensiones y realidades sobre la vulneración de los derechos fundamentales a falta de regulación de la inteligencia artificial (IA) en Colombia*. En: Nuevas tecnologías y el derecho, 45–62."

Asís, R. de. (2023). Inteligencia artificial y derechos humanos. En F. L. Ibáñez López-Pozas (Coord.), *Inteligencia artificial: los derechos humanos en el centro* (pp. 19-31). Dykinson.

Balaguer Callejón, F. (2025). Perspectivas metodológicas para el análisis constitucional de la inteligencia artificial y de su incidencia sobre los derechos fundamentales. En F. Balaguer Callejón & L. Escajedo San-Epifanio (Dirs.), *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales*. Dykinson.

Balaguer Callejón, F., & Escajedo San-Epifanio, L. (Dirs.). (2025). *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales*. Dykinson. DOI: https://doi.org/10.14679/4208

Barco Moreno, C. D., Mendoza Munar, L., & Urbano, M. K. (2023). *Inteligencia artificial, ética y regulación jurídica: Una mirada desde el derecho constitucional colombiano*. Universidad Cooperativa de Colombia.

Barrios Tao, H. (2025). Configuración de subjetividades mediante sistemas de inteligencia artificial. *Sophia, Colección de Filosofía de la Educación*, (39), 83–114. https://doi.org/10.17163/soph.n39.2025.02

Bekker, S. (2021). Fundamental rights in digital welfare states: The case of SyRI in the Netherlands. En O. Spijkers, W. G. Werner, & R. A. Wessel (Eds.), Netherlands Yearbook of International Law 2019 (Vol. 50, pp. 289–307). T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-403-7_24

Big Brother Watch. (2018). *Face Off: The lawless growth of facial recognition in UK policing*.

Big Brother Watch. (2020). *Briefing on facial recognition surveillance* (June 2020)

Cancio, R. C. (2023). Inteligencia artificial y administración de justicia: Una disrupción relativa. En F. L. Ibáñez López-Pozas (Coord.), *Inteligencia artificial: Los derechos humanos en el centro* (pp. 181–202). Dykinson.

Castro Serrano, Francisco de Borja, & Garay Rivera, José Miguel. (2025). Capturas y rupturas en los dispositivos de poder de la educación socioemocional. *Sophia, Colección de filosofía de la Educación*, (39), pp. 229-256.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven,

J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics, 3*, 241–255. https://doi.org/10.1007/s43681-022-00167-3

Chang, X. (2023). *Gender bias in hiring: An analysis of the impact of Amazon's recruiting algorithm*. In *Proceedings of the 2023 International Conference on Management Research and Economic Development* (pp. 134–140). https://doi.org/10.54254/2754-1169/23/20230367

Clavijo Cáceres, D., Guerra Moreno, D., & Yáñez Meza, D. (2014). *Método, metodología y técnicas de la investigación aplicada al derecho*. Grupo Editorial Ibáñez.

Coddou Mc Manus, A., Ortiz, M. G., & Tabares Soto, R. (2025). Evitando la trampa del formalismo: Evaluación crítica y selección de métricas de equidad estadística en algoritmos públicos. *Revista de Estudios Sociales, 93,* 85–106. https://doi.org/10.7440/res93.2025.05

Consejo Nacional de Política Económica y Social. (2025, 14 de febrero). *Política Nacional de Inteligencia Artificial* (Documento CONPES 4144). Departamento Nacional de Planeación.

Cuartielles Saura, R., & Carral, U. (2025). Herramientas de inteligencia artificial contra la desinformación populista. En F. Guerrero-Solé & L. Pérez-Altable (Eds.), *La democracia en riesgo: ¿Internet e IA al servicio de los populismos?* (pp. 115–121). Editorial UOC.

Dastin, J. (2018, 10 de octubre). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-

automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davidovic, J. (2023). On the purpose of meaningful human control of AI. *Frontiers in Big Data, 5*, 1017677. https://doi.org/10.3389/fdata.2022.1017677

Davies, B., Innes, M., & Dawson, A. (2018, septiembre). An evaluation of South Wales Police's use of automated facial recognition. Universities' Police Science Institute; Crime and Security Research Institute; Cardiff University

De Salvador Carrasco, L. (2023). Aprendizaje automático y protección de datos. En F. L. Ibáñez López-Pozas (Coord.), *Inteligencia artificial: los derechos humanos en el centro* (pp. 131–153). Dykinson.

Departamento Administrativo Nacional de Estadística. (2025). *Encuesta de Tecnologías de la Información y las Comunicaciones en Hogares (ENTIC Hogares) – 2024. Boletín técnico*. DANE.

Departamento Nacional de Planeación. (2023). *Estrategia Nacional Digital de Colombia 2023–2026*. DNP.

Díaz-Rincón, S. V., Enamorado-Estrada, J. A., y Bolaño-Retamozo, N. M. (2024). La inteligencia artificial en la solicitud e imposición de la medida de aseguramiento en el derecho penal de Colombia. Revista Jurídicas, 21(2), 199-220. https://doi.org/10.17151/jurid.2024.21.2.11.

Dieterich, W., Mendoza, C., y Brennan, T. (2016, 8 de julio). COMPAS risk scales: Demonstrating accuracy equity and predictive parity (Technical Report). Northpointe Inc.

Dressel, J., y Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Duque, S. (2025). La revolución silenciosa: inteligencia artificial, justicia social y cambio comunitario. *Naturaleza y Sociedad. Desafíos Medioambientales, 13*. https://doi.org/10.53010/nys.dia.09

Equivant. (2017). Practitioner's guide to COMPAS Core.

Escajedo San-Epifanio, L. (2025). Constitucionalismo y panóptico digital 'amable': Biometrías automatizadas, privacidad y libertad cognitiva. En F. Balaguer Callejón & L. Escajedo San-Epifanio (Dirs.), *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales* (pp. 69–105). Dykinson.

Fair Trials. (2024). Inteligencia artificial en la seguridad pública y en el sistema penal en América Latina: Análisis basado en el debido proceso. https://www.fairtrials.org.

Fiscalía General de la Nación. (2019). Herramienta PRiSMA: Perfil de Riesgo de Reincidencia para la Solicitud de Medidas de Aseguramiento. Dirección de Políticas Públicas y Estrategia.

Fiscalía General de la Nación. (2019). Informe de Avances Fiscalía General de la Nación 2016-2019. https://www.fiscalia.gov.co.

Fiscalía General de la Nación. (2019). La Fiscalía de la Gente. Periodo del Fiscal General de la Nación, Néstor Humberto Martínez Neira. 2016-2019 (Serie Documentos No. 34).

Fiscalía General de la Nación. (2020). Informe de empalme de la Fiscalía General de la Nación, 2020.

Fiscalía General de la Nación. (2025). Informe de Gestión 2024-2025.

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." Federal Probation, 80(2), 38–46.

Flórez Rojas, M. L. (2025). Algocracy in the judiciary: Challenging trust in the system. *Revista de Estudios Sociales, 93,* 107–128. https://doi.org/10.7440/res93.2025.06

Flórez Rojas, M. L. y Vargas Leal, J. (2020). El impacto de herramientas de inteligencia artificial: un análisis en el sector público en Colombia. En C. Aguerre (Ed.), Inteligencia Artificial en América Latina y el Caribe: Ética, Gobernanza y Políticas. CETyS Universidad de San Andrés.

Foucault, M. (1979). *Microfísica del poder* (J. Varela & F. Álvarez-Uría, Trads.). Las Ediciones de La Piqueta.

Foucault, M. (2002). *Vigilar y castigar: Nacimiento de la prisión* (A. Garzón del Camino, Trad.). Siglo XXI Editores. (Obra original publicada en 1975).

Galán Juárez, M. (2023). La Inteligencia Artificial en el contexto ético europeo: la autonomía humana. En F. L. Ibáñez López-Pozas (Coord.), *Inteligencia artificial: los derechos humanos en el centro* (pp. 49–58). Dykinson.

Galera Victoria, A. (2025). Inteligencia artificial, biometrías y tutela de derechos a la luz del Convenio Marco del Consejo de Europa y la jurisprudencia del TEDH. En F. Balaguer Callejón & L. Escajedo San-Epifanio (Dirs.), *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales* (pp. 269–297). Dykinson.

Garland, D. (1999). *Castigo y sociedad moderna: Un estudio de teoría social*. Siglo XXI Editores. (Obra original publicada en 1990).

Garland, D. (2005). *La cultura del control: Crimen y orden social en la sociedad contemporánea* (M. Sozzo, Trad.). Gedisa. (Obra original publicada en 2001).

Garland, D. (2019). Avances teóricos y problemas en la sociología del castigo [Theoretical advances and problems in the sociology of punishment]. *Delito y Sociedad, 28*(48), 9–30.

Gómez Pavajeau, C. A., Gómez Barranco, M. M., & Giraldo Velásquez, C. D. (2025). *La inteligencia artificial y la computación cuántica: Desafíos para el Derecho y la ética*. Grupo Editorial Ibáñez.

Gutiérrez, J. D., y Muñoz-Cadena, S. (2023). Adopción de sistemas de decisión automatizada en el sector público: Cartografía de 113 sistemas en Colombia. GIGAPP Estudios Working Papers, 10(270), 365-395.

Han, B.-C. (2022). *Infocracia: Digitalización y democracia*. Taurus.

Harari, Y. N. (2024). *Nexus: Una breve historia de la información*. Debate.

Harvard Law Review. (2017). Wisconsin Supreme Court requires warning before use of algorithmic risk assessments in sentencing. Harvard Law Review, 130(5), 1530–1537.

Hernández García de Velazco, J. J., Rhenals Turriago, J. E., & Álvarez Pertuz, A. A. (2025). Reformulaciones a la Teoría de los Derechos Humanos y Fundamentales a partir de la Inteligencia Artificial. Via Inveniendi Et Iudicandi, 20(1), 54-67. https://doi.org/10.15332/19090528.11113

Hernández Terán, M. (2025). *Neuroderechos, sesgos, daños, inteligencia artificial y otros problemas jurídicos*. Grupo Editorial Ibáñez.

Hernández-Sampieri, R., & Mendoza Torres, C. P. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill Interamericana Editores.

Herrera Giraldo, M. F., Gallego Acevedo, J. M., Gutiérrez Ramírez, L. H., Vargas, F., & Pereira, M. (2024). *La difusión de la inteligencia artificial en una economía emergente: Evidencia a nivel de la empresa en Colombia* (Nota técnica del BID No. IDB-TN-3067). Banco Interamericano de Desarrollo.

Ibáñez López Pozas, F. L. (2023). Experiencias y aplicaciones de la inteligencia artificial en la investigación criminal en el derecho comparado y su posible trasposición al derecho español. En F. L. Ibáñez López-Pozas (coord.), *Inteligencia artificial: los derechos humanos en el centro* (pp. 249-257). Dykinson.

Information Commissioner's Office. (2021). *The use of live facial recognition technology in public places* (Commissioner's Opinion).

Innerarity, D. (2025). *Una teoría crítica de la inteligencia artificial*. Galaxia Gutenberg.

Jiménez Martínez, M. V. (2025). La responsabilidad civil derivada de los efectos discriminatorios en el perfilado de las personas. En F. Balaguer Callejón & L. Escajedo San-Epifanio (Dirs.), *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales* (pp. 467–502). Dykinson.

Lazcoz Moratinos, G., & Castillo Parrilla, J. A. (2020). Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: El caso SyRI. Revista Chilena de Derecho y Tecnología, 9(1), 207–225. https://doi.org/10.5354/0719-2584.2020.56843

Leal Martínez, A. M. (2021). ¿Se puede marginar y resocializar al mismo tiempo? Apuntes para un modelo de rehabilitación penitenciaria en Colombia. En L. J. Ariza, M. Iturralde, & F. L. Tamayo Arboleda (Eds.), *Cárcel, derecho y sociedad: Aproximaciones al mundo penitenciario en Colombia* (pp. 225–245). Universidad de los Andes.

López Vega, J. E., Becerra, V., Guzmán, M. A., & Landazuri Sandoval, J. K. (2023). *Inteligencia artificial en la justicia colombiana: ¿la solución a la congestión judicial?* Revista Lecciones Vitales, 1, Artículo lv0101. https://doi.org/10.18046/rlv.2023.5655

Mayson, S. G. (2018). Bias in, bias out. Yale Law Journal, 128, 2218–2300.

Medina Uribe, P. y Gómez, L. F. (26 de julio de 2020). Unificar las causas, agilizar los trámites y las dudas sobre su uso en el futuro: ¿cómo es y qué busca el software que utiliza la justicia colombiana? ColombiaCheck.

Meuwese, A. (2020). Regulating algorithmic decision-making one case at the time: A note on the Dutch 'SyRI' judgment. European Review of Digital Administration & Law, 1(1-2), 209–211. https://doi.org/10.4399/978882553896019

Ministerie van Sociale Zaken en Werkgelegenheid. (2014, 1 de septiembre). Besluit van 1 september 2014 tot wijziging van het Besluit SUWI in verband met regels voor fraudeaanpak door gegevensuitwisselingen en het effectief gebruik van binnen de overheid bekend zijnde gegevens met inzet van SyRI. Staatsblad van het Koninkrijk der Nederlanden, 2014(320).

Morales Higuita, L., Agudelo Londoño, S., Montoya Raigosa, M. y Montoya Vidales, A. M. (2021). Inteligencia artificial en el proceso penal: análisis a la luz del Fiscal Watson. Pensamiento Jurídico, (54), 147-164.

Moreno Blanco, N. (2019). *Inteligencia artificial y reincidencia: Control inteligente del delito. Aproximación al Sistema Prisma de la Fiscalía General de la Nación*. Universidad de los Andes.

Office of the United Nations High Commissioner for Human Rights. (2020, 5 de febrero). Landmark ruling by Dutch court stops government attempts to spy on the poor – UN expert [Comunicado de prensa]. https://www.ohchr.org/en/press-releases/2020/02/landmark-ruling-dutch-court-stops-government-attempts-spy-poor-un-expert

Palacios, L., Forero, V., & Labarthe, S. (2024). *Fiscal Watson: estudio del uso de inteligencia artificial en la Fiscalía General de la Nación en Colombia*. Derechos Digitales.

Palacios, L., Forero, V., & Labarthe, S. (2024). Fiscal Watson: estudio del uso de Inteligencia Artificial en la Fiscalía General de la Nación en Colombia. Derechos Digitales. https://ia.derechosdigitales.org/.

Pasquale, F. (2020). *New laws of robotics: Defending human expertise in the age of AI*. Harvard University Press.

Pawlik, M. (2022). *El deber de cooperación ciudadano en derecho penal y la posición de los excluidos*. Universidad Externado de Colombia.

Pérez Conchillo, E. (2025). Reflexiones conceptuales sobre la transparencia algorítmica aplicadas a las plataformas digitales para la protección de los derechos y la democracia. En F. Balaguer Callejón & L. Escajedo San-Epifanio (Dirs.), *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales* (pp. 437–461). Dykinson.

Pernas Ciudad, E. (2023). Inteligencia artificial e igualdad de género. En F. L. Ibáñez López-Pozas (Coord.), *Inteligencia artificial: Los derechos humanos en el centro* (pp. 99-110). Dykinson.

Peter Fussey, & Daragh Murray. (2019). *Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology*. Human Rights Centre, University of Essex.

Presno Linera, M. Á. (2025). La prohibición de los sistemas que evalúan y clasifican a las personas con efectos discriminatorios en el Reglamento Europeo de Inteligencia Artificial. En F. Balaguer Callejón & L. Escajedo San-Epifanio (Dirs.), *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales* (pp. 507-520). Dykinson.

Programa de las Naciones Unidas para el Desarrollo (PNUD). (2024). *AILA: Evaluación del panorama de la inteligencia artificial en Colombia*. PNUD.

Purshouse, J., & Campbell, L. (2020). Automated facial recognition and policing: A Bridge too far? University of East Anglia; Monash University.

Quiceno Osorio, J. D. (2025). La inteligencia artificial y el riesgo de una analogía invertida. *Sophia, Colección de Filosofía de la Educación*, (39), 315-335. https://doi.org/10.17163/soph.n39.2025.10

Rachovitsa, A., & Johann, N. (2022). The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. Human Rights Law Review, 22(2), 1–15. https://doi.org/10.1093/hrlr/ngac010

Rivero Ortega, Ricardo. (2023). Derecho e inteligencia artificial. Ediciones Olejnik.

Rodas Florián, D. P. R. (2024). *Análisis legal y jurisprudencial del impacto de la discriminación algorítmica en el derecho a la igualdad en la era de la inteligencia artificial*. Universidad Cooperativa de Colombia.

Romano, A. (2025). La inteligencia artificial emocional en el ámbito migratorio y sus riesgos para los derechos fundamentales: ¿hacia la biodeterminación de la credibilidad? En F. Balaguer Callejón y L. Escajedo San-Epifanio (Dirs.), *Vigilancia biométrica masiva, inteligencia artificial y derechos fundamentales* (pp. 561-580). Dykinson. https://doi.org/10.14679/4208.

Sánchez Barrilao, J. F. (2016). El Derecho constitucional ante la era de Ultrón: la informática y la inteligencia artificial como objeto constitucional. *Estudios de Deusto: Revista de Derecho Público*, *64*(2), 225-258. https://doi.org/10.18543/ed-64(2)-2016pp225-258

Sánchez Vásquez, C. (2021). *El derecho al control humano en la inteligencia artificial: Una propuesta de regulación del control humano como un nuevo derecho en el ordenamiento jurídico colombiano*. Universidad EAFIT.

Santana Ramos, E. M. (2025). El impacto de la inteligencia artificial en la construcción de la identidad y la autonomía personal. *Islas, 67*(211), e1598. Universidad Central "Marta Abreu" de Las Villas. https://islas.uclv.edu.cu/index.php/islas/article/view/1598

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI, 5*, 15. https://doi.org/10.3389/frobt.2018.00015

Sarrión Esteve, J. (2023). El derecho constitucional ante la inteligencia artificial: perspectivas y retos que plantean las tecnologías inteligentes. En F. L. Ibáñez López-Pozas (coord.), *Inteligencia artificial: los derechos humanos en el centro* (pp. 79-97). Dykinson.

Solar Cayón, J. I. (2020). La inteligencia artificial jurídica: nuevas herramientas y perspectivas metodológicas para el jurista. *Revus: Journal for Constitutional Theory and Philosophy of Law, 41*, 1–27. https://doi.org/10.4000/revus.6547

South Wales Police. (2025a, 22 de febrero). *Joint Data Protection Impact Assessment (DPIA) and Information Security Impact Assessment: Live Facial Recognition (LFR) & Op Gwalia*. Digital Services Division; South Wales Police

South Wales Police. (2025b). *Facial Recognition Technology: Equality Impact Assessment* (v1.4).

Surveillance Camera Commissioner. (2019). *Guidance for the police use of overt surveillance camera systems incorporating facial recognition technology to locate persons on a watchlist*.

Tirado Serrano, F. J., & Blasco Ejarque, J. L. (2025). Gubernamentalidad algorítmica, inteligencia artificial y prácticas de disidencia en Latam. *Psicoperspectivas, 24*(2). https://doi.org/10.5027/psicoperspectivas-Vol24-Issue2-fulltext-3441

Torres Abril, J. S., Silva Vásquez, Z. Y., & Gómez Simijaca, V. (2022). Herramientas de la inteligencia artificial dentro del sistema judicial colombiano: Estudio del caso PRETORIA y PRISMA. *Revista Principia Iuris, 19*(40), 48–60.

Torres Abril, J. S., Silva Vásquez, Z. Y., y Gomez Simijaca, V. (2022). Herramientas de la inteligencia artificial dentro del sistema judicial colombiano; estudio del caso PretorIA y PRISMA. Revista Principia Iuris, 19(40), 48-67.

Universidad de los Andes. (2024). *Perfil de riesgo de recurrencia de la solicitud de medidas penitenciarias (PRISMA)*. Sistemas de Algoritmos Públicos, Universidad de los Andes. https://algoritmos.uniandes.edu.co/perfil-de-riesgo-de-recurrencia-de-la-solicitud-de-medidas-penitenciarias-prisma-2/

Universidad de los Andes. (2025, 29 de enero). Perfil de Riesgo de Recurrencia de la Solicitud de Medidas Penitenciarias (Prisma). Sistemas de Algoritmos Públicos.

van Bekkum, M., & Zuiderveen Borgesius, F. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security, 23*(4), 323–340. https://doi.org/10.1177/13882627211031257 (SAGE Journals)

Van Bekkum, M., & Zuiderveen Borgesius, F. J. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. European Journal of Social Security, 23(4), 323–340. https://doi.org/10.1177/13882627211031257

Vázquez Pita, E. (2025). Evolución del concepto de inteligencia artificial (IA) y sus consecuencias jurídicas. *Revista de Derecho UNED, (35),* 437–456. https://doi.org/10.5944/rduned.35.2025.45884

Washington, A. L. (2018). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. Colorado Technology Law Journal, 17(1), 131–160.

Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. Data & Policy, 5, e2. https://doi.org/10.1017/dap.2022.39

Younes Moreno, D. (2021). *Derecho constitucional colombiano* (17.ª ed.). Legis.

Zuboff, S. (2019). *La era del capitalismo de la vigilancia*. Paidós.

Apéndice A.

Glosario

Human-in-the-loop: este término refiere a una práctica de verificación que integra la inteligencia artificial conservando la intervención humana central. Se describe como un «procedimiento parcial, puesto que el sistema requiere de una base humana», cuyo criterio y juicio conservan su «posición central (human-in-the-loop)». Esta aproximación es fundamental en el llamado «fact-checking computacional» o asistido, asegurando que la tecnología no opere de forma aislada. La metodología busca que los resultados generados automáticamente sean validados por el discernimiento de una persona, garantizando que el proceso de comprobación de hechos mantenga un estándar humano (Cuartielles Saura & Carral, 2025, p. 117).

Métricas de equidad: la equidad estadística (*statistical fairness*) consiste en «medir la equidad a partir de criterios objetivos, de métricas que permitan decir si un determinado sistema automatizado cumple con cierto estándar de justicia previamente establecido». Estas métricas estadísticamente centradas «permiten medir la equidad de los resultados en términos de distribución entre diferentes grupos poblacionales». Ejemplos comunes incluyen la «paridad demográfica», que busca asegurar que la proporción de resultados positivos sea igual para todos los grupos, o el impacto dispar, que evalúa si decisiones aparentemente neutrales afectan negativamente a grupos protegidos por la ley (Coddou et al., 2025).

Regla de inferencia: la regla de inferencia es un componente de los sistemas expertos que «despliega patrones de razonamiento y búsqueda a lo largo de la base de conocimiento para encontrar soluciones». En el diseño tradicional, era imperativo «codificar

conocimientos y reglas de inferencia» de forma manual por especialistas. No obstante, en la inteligencia artificial actual, estas han sido sustituidas por un «esquema de funcionamiento analógico que detecta correlaciones» entre múltiples elementos. El sistema ya no depende de premisas predeterminadas, sino que infiere automáticamente las reglas de asociación aprovechando la capacidad de procesamiento de grandes volúmenes de datos (Solar, 2020).

Proxies sensibles: los proxies sensibles son variables aparentemente neutrales que guardan una correlación estrecha con atributos protegidos, induciendo discriminación indirecta. Un modelo puede tener un «impacto dispar si las variables que emplea, como el código postal o la ocupación, están correlacionadas con la raza». En estos casos, la variable neutral actúa como un proxy del dato sensible. Para detectar estos sesgos, es posible elaborar «datos sintéticos que simulen escenarios hipotéticos donde se modifiquen las características sensibles de los individuos». Esto permite evaluar si el algoritmo mantiene la equidad al invertir factores como el género en el perfil analizado (Coddou et al., 2025).

Proxies: los proxies son indicadores «superficiales pero efectivos» que los algoritmos emplean para identificar categorías mediante correlaciones estadísticas. Al no considerar el caso individual, los sistemas utilizan estos sustitutos para deducir información oculta o sensible; por ejemplo, el autor explica que «el código postal permite deducir la raza». Este fenómeno facilita clasificar personas a través de huellas indirectas, dado que las etiquetas discriminatorias a menudo no son explícitas, sino que «obedecen al fenómeno proxy» (Innerarity, 2025).